A version of bundle method with linear programming

Shuai Liu,[*] Andrew Eberhard,[†] and Yousong Luo[‡]

School of Mathematical and Geospatial Sciences,
RMIT University, Melbourne, VIC 3001, Australia

July 7, 2015

Abstract

Bundle methods have been intensively studied for solving both convex and nonconvex optimization problems. In most of the bundle methods developed thus far, at least one quadratic programming (QP) subproblem needs to be solved in each iteration. In this paper, we exploit the feasibility of developing a bundle algorithm that only solves linear subproblems. We start from minimization of a convex function and show that the sequence of major iterations converge to a minimizer. For nonconvex functions we consider functions that are locally Lipschitz continuous and prox-regular on a bounded level set, and minimize the cutting-plane model over a trust region with infinity norm. The para-convexity of such functions allows us to use the locally convexified model and its convexity properties. Under some conditions and assumptions, we study the convergence of the proposed algorithm through the outer semicontinuity of the proximal mapping. Encouraging results of preliminary numerical experiments on standard test sets are provided.

Keywords Nonconvex optimization, Nonsmooth optimization, Trust region method, Linear subproblem, Prox-regular

1 Introduction

As a generalization of nonlinear programming (NLP), nonsmooth optimization (NSO) has broader application areas and more theoretical challenges. Traced back to as early as 1962 [24], NSO has been studied intensively hitherto with more and more new methods being developed. Generally speaking, methods in unconstrained NSO fall into the following categories: subgradient methods [25, 12, 1], gradient sampling methods [2, 26], modified Newton [20] or quasi-Newton [16] methods, proximal point methods [21], derivative free methods [27], and bundle methods [15, 28, 14, 23]. NSO has a much broader scope than NLP and theory can be developed for very generic functions classes ie. Lipschitz continuous etc. So in order to develop more efficient algorithms and stronger theoretic results, various special structures of the objective function are often assumed. Special algorithms have developed, tailored for well-structured problems such as convex composite [10], partially separable [19], etc. Bundle method is probably the most intensively researched methods for nonsmooth optimization. It was created independently by Claude Lemaréchal [15] and Philip Wolfe [28] in 1975. Since then a great number of variants of bundle methods have been developed, such as proximal bundle [14, 6], trust region bundle [23, 4], splitting bundle [7], and redistributed bundle [8]. Bundle methods grew out of cutting plane methods which often showed a great deal of instability (see page 276 of [9]). To correct this a better approximation of the whole subdifferential is made and stabilization of the descent step is also incorporated.

Consider any optimization problems of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f(x) \tag{1}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous. The traditional bundle method solves a parametric quadratic subproblem in each iteration to obtain a search direction with possible options to

follow with a line search. The advantage of employing a quadratic subproblem is that the optimal solution is unique due to the strict convexity, furthermore, the solution can be explicitly expressed in terms of the current iteration point and an average of some subgradients. However this convenience must be weighed up against the inconvenience of having to solve a QP at each iteration which can be time-consuming, especially for large scale problems. One of the motivations of this paper is to overcome the necessity of solving a QP in each iteration. Hence, we propose a new version of bundle algorithm for solving (1) with linear subproblems only. Surprisingly, one finds that even without the explicit expression of the solution, and even without line search, our algorithm can still converge to stationary points for nonconvex problems.

Our approach uses two key tools to achieve this end. First, is a subdifferential approximation. Traditional bundle methods use a convex combination of subgradients (an average) to approximate a selection from an approximation of the whole subdifferential set (the subgradient selection technique). Traditional subgradient methods, including gradient sampling methods, use minimum norm subgradient as a replacement for the gradient, a calculation involving the solution of a QP. In our method we use a trust region method based on a linear cutting plane model (of a related local convexification of f) and in the theoretical analysis we use $\frac{f(x)-\bar{f}}{\|x-P(x)\|_\infty}$, where $\bar{f}$ is the minimum value of f and $P(x)$ is the projection of x onto the optimal solution set, as a lower bound on the norm of the minimum norm element of the subdifferential. This alternative for the subdifferential approximation plays a significant role in our convergence analysis. Secondly, we use a local convexification technique developed in [8]. For nonconvex functions that are prox-regular and Lipschitz continuous, we show that there exists a number 'a' such that $f(\cdot) + \frac{a}{2}\|\cdot - \bar{x}\|^2$ is a restriction, to a level set of f, of a globally convex function. Clearly when f is convex it suffices to take $a = 0$. Unlike [8], where the convexification parameter 'a' is eventually stabilized, we only need 'a' to be bounded. This allows us to exploit the outer semicontinuity of the subdifferential and that of the associated *proximal mapping*, while traditional bundle methods use the outer semicontinuity of the ϵ-subdifferential.

The use of a linear model in a trust region method is a relatively new idea and our work follows that of [17] where a related approach is used to solve a large scale optimization problem that arise out of stochastic programming. Standard theory for trust region methods use quadratic model functions. This is usually justified via the imposition of sufficient differentiability assumption on the objective function. Such logic is less compelling in the context of nonsmooth optimization. Furthermore, one usually associates a bundle–trust–region method to an approach that absorbs the trust region into a quadratic penalty to control the step length. We instead directly impose an infinity box norm that is handled in more or less a tradition manner for a trust region method. Unlike [17] we are able to handle a generic class of prox-regular, locally Lipschitz function, greatly extending the applicability of this approach.

This paper is organized as follows. Section 2 contains the properties of our objective function as preliminary knowledge and Section 3 includes a description of a version of bundle method with linear subproblem for convex optimization. In Section 4 we derive the method for nonconvex optimization and we analyse the convergence of the algorithm in Section 5. Preliminary numerical tests are presented in Section 6.

In this paper, $\|\cdot\|$, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote the two norm, one norm and infinity norm, respectively. Denote by $\mathrm{lev}_b f$ the lower level set of f defined by $\{x \in \mathbb{R}^n | f(x) \leq b\}$ and $B(x, \epsilon)$ the closed ball centered at x with radius ϵ. The convex hull of a set $C \in \mathbb{R}^n$ is denoted by $\mathrm{co}\, C$, the domain of function f is $\mathrm{dom}\, f$ and the interior of a set C is $\mathrm{int}\, C$.

2 Properties of the objective function

For the reader's convenience we collect in this section some standard definitions and properties we will be utilizing in our development.

Definition 2.1 (subdifferential). Let $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ be finite at $\bar{x}$.

1. The set

$$\hat{\partial} f(\bar{x}) := \left\{ s \in \mathbb{R}^n \,\Big|\, \liminf_{\substack{x \to \bar{x} \\ x \neq \bar{x}}} \frac{f(x) - f(\bar{x}) - \langle s, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0 \right\}$$

is called the *Fréchet subdifferential* or *regular subdifferential* of f at $\bar{x}$ with elements called *Fréchet subgradients* or *regular subgradients* of f at $\bar{x}$. When $|f(\bar{x})| = \infty$ then $\hat{\partial} f(\bar{x}) := \emptyset$.

2. The set $\partial f(\bar{x}) := \limsup\limits_{\substack{x \to \bar{x} \\ f(x) \to f(\bar{x})}} \hat{\partial} f(\bar{x})$ is called the *basic subdifferential* of f at $\bar{x}$ with elements called *basic subgradients* of f at $\bar{x}$. When $|f(\bar{x})| = \infty$ then $\partial f(\bar{x}) := \emptyset$.

Definition 2.2 (prox-regularity, Definition 13.27, [22]). A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *prox-regular* at $\bar{x}$ for $\bar{v}$ with respect to ϵ and a if f is finite and locally lower semicontinuous (l.s.c) at $\bar{x}$ with $\bar{v} \in \partial f(\bar{x})$, and there exist $\epsilon > 0$ and $a \geq 0$ such that

$$f(x') \geq f(x) + \langle v, x' - x \rangle - \frac{a}{2}\|x' - x\|^2 \ \forall \ x' \in B(\bar{x}, \epsilon) \tag{2}$$

when $\|x - \bar{x}\| < \epsilon$, $v \in \partial f(x)$, $\|v - \bar{v}\| < \epsilon$, $f(x) < f(\bar{x}) + \epsilon$. When this holds for all $\bar{v} \in \partial f(\bar{x})$, f is said to be prox-regular at $\bar{x}$.

Definition 2.3 (para-convexity). Given a point $\bar{x} \in \mathbb{R}^n$ and a real number $\epsilon > 0$, a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *para-convex* on $B(\bar{x}, \epsilon)$ with respect to a if there exists $a \geq 0$ such that the function $f(\cdot) + \frac{a}{2}\|\cdot\|^2$ is convex on $B(\bar{x}, \epsilon)$.

Definition 2.4 (proximal mapping). For a proper, l.s.c. function $f : \mathbb{R}^n \mapsto \bar{\mathbb{R}}$ and a parameter $a > 0$, the *Moreau envelope function* $e_a f$ and *proximal mapping* (or *proximal point mapping*) $P_a f$ are defined by

$$e_a f(x) := \inf\limits_{w \in \text{dom} f}\{f(w) + \frac{a}{2}\|w - x\|^2\}, \tag{3}$$

$$P_a f(x) := \arg\min\limits_{w \in \text{dom} f}\{f(w) + \frac{a}{2}\|w - x\|^2\}. \tag{4}$$

If the proper l.s.c function f is bounded from below then $P_a f(x)$ is nonempty and compact for all $(x, a) \in \mathbb{R}^n \times \mathbb{R}_{>0}$, and the mapping $\mathbb{R}^n \times \mathbb{R}_{>0} \ni (x, a) \mapsto e_a f(x)$ is continuous. From Definition 2.4 we see that if $p \in P_a f(x)$ then $f(p) \leq e_a f(x)$ and $a(p - x) \in \partial f(p)$. We also have $e_a f(x) \leq f(x)$ for all $x \in \mathbb{R}^n$ and $e_a f(x) = f(x)$ if and only if $x \in P_a f(x)$.

Definition 2.5 (outer semicontinuity). A set valued mapping $S : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ is *outer semicontinuous* at $\bar{x}$ if

$$\{u \mid \exists \ x^j \to \bar{x}, \ \exists \ u^j \to u \text{ with } u^j \in S(x^j)\} = S(\bar{x}).$$

We note that both subdifferential and proximal mapping are outer semicontinuous. Additionally, the mapping $\mathbb{R}^n \times \mathbb{R}_{>0} \ni (x, a) \mapsto P_a f(x)$ is outer semicontinuous.

The following proposition is from lemma 2.2 of [5].

Proposition 1. *If a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is locally Lipschitz continuous and prox-regular at $\bar{x}$ then there exist ϵ and 'a' such that f is para-convex on $B(\bar{x}, \epsilon)$ with respect to 'a'.*

Remark. If $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex function then it is useful to note that

$$f(x) = \sup\left\{f(y) + s^T(x - y) \mid y \in \mathbb{R}^n, \ s \in \partial f(y)\right\}. \tag{5}$$

3 LP-bundle method : The convex case

Consider minimizing a convex function f on $\mathbb{R}^n$. Denote S the set of minimizers of f. Then S is closed and convex. Assuming S is nonempty, the projection operator $P(\cdot)$ onto S is well defined. In reference [17], a bundle trust-region method was proposed to solve a two-stage stochastic linear programming problem. We show that this method can be generalized to minimize any convex and locally Lipschitz continuous functions. We refer to the generalized method as bundle method with linear programming for convex optimization (LPBC). Given an auxiliary point y_i and a subgradient $s_i \in \partial f(y_i)$, a cutting-plane function is a linear mapping

$$x \mapsto f(y_i) + \langle s_i, x - y_i \rangle. \tag{6}$$

The cutting-plane model of $f(x)$ is constructed by the point-wise maximum of the cutting-plane functions as follows:

$$m(x) = \max\limits_{i \in I}\{f(y_i) + \langle s_i, x - y_i \rangle\}, \tag{7}$$

3

where I is the index set of auxiliary points. LPBC applies the model function on a trust region generated by infinity norm so that it solves the following subproblem sequentially,

$$\min_{x \in \mathbb{R}^n} \quad m(x)$$
$$\text{subject to} \quad ||x - \bar{x}||_\infty \leq \Delta, \tag{8}$$

where $\bar{x}$ is the current best candidate for a minimizer of f and Δ is the trust region radius. Adopting a scalar variable, problem (8) is equivalent to the following linear programming problem

$$\min_{(x,z) \in \mathbb{R}^{n+1}} \quad z \tag{9a}$$
$$\text{subject to} \quad f(y_i) + \langle s_i, x - y_i \rangle \leq z, \ \forall \, i \in I, \tag{9b}$$
$$||x - \bar{x}||_\infty \leq \Delta. \tag{9c}$$

During the kth iteration, LPBC solves several linear problems with different model functions 'm' and possibly different trust region radii Δ before a new iterate x^{k+1} is identified. Hence LPBC refers to x^k and x^{k+1} as *major iterates* and x^{kl}, $l = 0, 1, 2, \cdots$ obtained by solving the current linear problem as *minor iterates*. We will also use x^* to denote minor iterates when it is not necessary to identify the iteration indices. The subscript kl and sometimes (k,l) means l minor iterations have been executed after k-th major iteration. Consequently, the $\bar{x}$, Δ and I in (9) are replaced by x^k, Δ_l^k and $I(k,l)$. After solving subproblem (9), an optimal solution (x^{kl}, z^{kl}) is obtained. And x^{kl} will be accepted as new iterate x^{k+1} if it yields substantial reduction in the real objective f, otherwise the model function will be refined by adding and deleting cutting planes. The substantial reduction in the value of f is measured by its quotient with the reduction of model value, i.e. $m_l^k(x^k) - m_l^k(x^{kl})$. LPBC updates the model m in a way such that the following conditions hold:

$$m_l^k(x^k) = f(x^k), \quad \text{for all index } k, \ l. \tag{10}$$

$$m_l^k \text{ is a convex, piecewise linear lower underestimate of } f, \ l = 1, 2, \cdots. \tag{11}$$

Specifically, to obtain m_{l+1}^k, LPBC flushes all the cutting planes except the following two types.

- The cutting plane is generated at x^k. Thus the cutting plane $f(x^k) + s^{k^T}(x - x^k)$ is always kept in the linear subproblem during the kth major iteration;

- The cutting plane is active at x^{kl} with positive Lagrange multiplier.

LPBC adds the new cutting plane generated at x^{kl}, $f(x^{kl}) + s_l^{k^T}(x - x^{kl})$ to the model m_{l+1}^k.

Procedure 1: LPBC Updating Trust Region

Define
$$\rho_l^k := \frac{f(x^k) - f(x^{kl})}{f(x^k) - m_l^k(x^{kl})} \tag{12}$$

if $\rho_l^k > \eta_3$ **and** $||x^{kl} - x^k||_\infty > 0.9\Delta$ **then**
   | $\Delta \leftarrow \min\{\alpha_2\Delta, \Delta_{\max}\}$
else if $\rho_l^k < -\frac{1}{\min\{1,\Delta\}}$ **then**
   | $\Delta \leftarrow \alpha_1\Delta$;
end

In the definition of ρ_l^k in (12), the denominator $f(x^k) - m_l^k(x^{kl})$ is the reduction of model $m(x)$ from x^k to x^{kl} due to condition (10). We will use the notion *linearization error*, the difference between the value of a function and the value of a cutting-plane function. Consider a cutting plane of a generic function, as defined in (6). The linearization error of this cutting plane at $\bar{x}$ is

$$\tilde{e}_i := f(\bar{x}) - [f(y_i) + \langle s_i, \bar{x} - y_i \rangle]. \tag{13}$$

In the LPBC algorithm, the linear subproblem (9) was used. Applying the KKT condition to (9), we can deduce the explicit expression of the model reduction of LPBC as following,

$$
m(\bar{x}) - m(x^*) = \begin{cases} \sum\limits_{i \in \bar{I}} \lambda_i \tilde{e}_i, & \text{if } \|x^* - \bar{x}\|_\infty < \Delta; \\ \sum\limits_{i \in \bar{I}} \lambda_i \tilde{e}_i + \Delta \| \sum\limits_{i \in \bar{I}} \lambda_i s_i \|_1, & \text{if } \|x^* - \bar{x}\|_\infty = \Delta. \end{cases} \tag{14}
$$

where $\bar{I}$ is the index set for active constraints in (9b) associated with an optimal solution (x^*, z^*), and λ_i for $i \in \bar{I}$ are the corresponding Lagrangian multipliers associated with (x^*, z^*). Since f is convex, all the linearization errors $\tilde{e}_i$ will be nonnegative. We also have

$$
\sum_{i \in \bar{I}} \lambda_i s_i \in \partial_{\bar{\epsilon}} f(\bar{x}), \text{ where } \bar{\epsilon} = \sum_{i \in \bar{I}} \lambda_i \tilde{e}_i. \tag{15}
$$

The derivation of (14) and (15) can be found in Lemma 2 where we prove the same conclusion for the generalization of LPBC, the LPBNC algorithm, with derivation following the same reasoning. The mapping $(x, \epsilon) \mapsto \partial_\epsilon f(x)$ is outer-semicontinuous, and hence when the model reduction decreases to 0 we have $0 \in \partial f(x)$. Consequently, our stopping criterion is that the model reduction is sufficiently small as it is showed in line 3 of Algorithm 2. It is worthy to note that our model reduction in (14) is comparable with that in the classical bundle method which uses a quadratic model of the form

$$
\min_{x \in \mathbb{R}^n} \left[m(x) + \frac{\mu}{2} \|x - \bar{x}\|^2 \right]. \tag{16}
$$

If the above model is used, then $m(\bar{x}) - m(x^*) = \sum\limits_{i \in \bar{I}} \lambda_i \tilde{e}_i + \frac{1}{\mu} \| \sum\limits_{i \in \bar{I}} \lambda_i s_i \|^2$. A significant difference between (8) and (16) is that the latter is strictly convex but the former is not. The optimal solution to (16) is unique and can be expressed by $x^* = \bar{x} - \frac{1}{\mu} \sum\limits_{i \in \bar{I}} \lambda_i s_i$, while (8) may have multiple solutions. The readers are referred to chapter XIV and XV of [9] for a comprehensive understanding of classical bundle methods. Algorithm LPBNC

Algorithm 2: Algorithm LPBC

Data: Final accuracy tolerance ϵ_{tol}, and maximum trust region radius $\Delta_{\max}$, initial trust region $\Delta_1^1 \in [1, \Delta_{\max})$, initial point x^0, trust region parameters η_1, η_2, η_3, integer $T \geq 20$ (inactive threshold);

1 *Initialization* Set the major and minor iteration counter $(k, l) \leftarrow (0, 0)$;
2 generate a cutting plane at x^k, update the index set $I(k, l)$ and define the cutting plane model (7). Set the minor iteration counter $l = 0$, $y_1^k = x^k$ and compute $s_1^k \in \partial f(y_1^k)$;
3 Solve the linear programming subproblem (9) and obtain an optimal solution (x^{kl}, z^{kl});
4 **if** $f(x^k) - m_l^k(x^{kl}) \leq (1 + |f(x^k)|)\epsilon_{\text{tol}}$ **then**
5 $\quad$ STOP
6 **end**
7 **if** $\rho_l^k = \frac{f(x^k) - f(x^{kl})}{f(x^k) - m_l^k(x^{kl})} \geq \eta_1$ **then**
8 $\quad$ $x^{k+1} = x^{kl}$;
9 $\quad$ obtain Δ_0^{k+1} via procedure 1;
10 $\quad$ obtain m_0^{k+1} by keeping all cutting planes except those that have been inactive for T iterations;
11 $\quad$ $k = k + 1$, continue to next major iteration by going to line 2
12 **else**
13 $\quad$ obtain Δ_{l+1}^k via procedure 1;
14 $\quad$ delete all cutting planes except the one generated at x^k, those that are active at x^{kl} and those that have been inactive for less than T times.
15 **end**
16 add the cutting plane $f(x^{kl}) + s_l^{k T}(x - x^{kl})$ to the model m_{l+1}^k;
17 set $l = l + 1$ and go to line 2;

---

is a generalization of LPBC and we will provide a convergence proof for this in section 5. Hence we omit the proof here and only state the following convergence theorem for the convex case.

Theorem 1. *Suppose that $\epsilon_{\text{tol}} = 0$.*
(i) If Algorithm 2 terminates at x^{kl}, then x^k is a minimizer of f with $x^k = P(x^k)$;
(ii) if there is an infinite number of minor iterations during the kth major iteration, then x^k is a minimizer of f with $x^k = P(x^k)$ and $\lim_{l\to\infty} m_l^k(x^{kl}) - f(x^k) = 0$;
(iii) if the sequence of major iterations $\{x^k\}$ is infinite then $\lim_{k\to\infty} \|x^k - P(x^k)\|_\infty = 0$.

4   LP-bundle method : The nonconvex case

4.1   Derivation of the Method

Based on the LPBC, we can derive a nonconvex version of the method through convexification for special types of functions. Specifically, we consider locally Lipschitz continuous functions that are prox-regular. Such functions are para-convex. Hence we can use a linear model to approximate a locally convex function. Under some assumptions, we show that the accumulation point of the minimizers of such functions is a stationary point of the objective function via the theory of proximal point mapping.

First we state the assumption on the objective function.

Assumption 1. *The objective function f is locally Lipschitz continuous and bounded below. Given $x^0 \in \mathbb{R}^n$ f is prox-regular on bounded level set $\text{lev}_{x^0} f$.*

Note that a single-valued function f is prox-regular on an open set O and locally Lipschitz continuous is equivalent to that f is lower-C^2 on O; see [22, Prop. 13.33]. Define

$$g(y) := g(y; x, a) : y \mapsto f(y) + \frac{a}{2}\|y - x\|^2 \tag{17}$$

with x and a as parameters. Under Assumption 1, the Moreau envelope function and the proximal point mapping associated with the objective f are globally well defined. We redefine them here as

$$e_a(x) := \min_{y\in\mathbb{R}^n}\{g(y; x, a)\} \quad P_a(x) := \arg\min_{y\in\mathbb{R}^n}\{g(y; x, a)\}. \tag{18}$$

If f is para-convex, then according to Definition 2.3, $g(y)$ is convex on some neighborhood $B_b(x)$. Clearly, for different x and b, in order to make $g(y)$ convex with respect to y, there exists a threshold for the value of 'a'. The motivation of our method is based on the following observation. Suppose we have some sequences $x^k \to x'$, $a^k \to a'$, and $b^k \to b'$ such that $g(y; x^k, a^k)$ is convex with respect to y on $B_{b^k}(x^k)$ for all k. Then we can use a cutting-plane model for $g(y; x^k, a^k)$ with trust region as in LPBC to obtain descent locally. To justify its stationarity, x' should be a global minimizer of $g(y; x', a')$. In fact, the outer semicontinuity of the mapping $(x, a) \mapsto P_a(x)$ means if there exist $x^k \to \bar{x}$, $\{a^k\}$ bounded and $p^k \in P_{a^k}(x^k)$ with $\|x^k - p^k\| \to 0$, then $\bar{x} \in P_{\bar{a}}(\bar{x})$ for some $\bar{a}$. Thus in order to find a stationary point our goal can be translated to generating a sequence $\{x^n\}$ and $\{p^n\}$ such that $\lim_{n\to\infty} \|x^n - p^n\|_\infty = 0$ with $p^n \in P_{a^n}(x^n)$ and $\{a^n\}$ bounded. Further discussion will be made in Section 5.

The following lemma shows that there exists a threshold for the value a such that the function $g(y)$ is locally convex on $\text{lev}_{x^0} f$ (which is not necessarily convex).

Lemma 1. *Under assumption 1, there exists a number $\bar{a} \geq 0$ such that for any $a \geq \bar{a}$, the function $g(y; x, a)$ is convex on a neighborhood of y for all $y \in \text{lev}_{x^0} f$.*

Proof. According to Assumption 1, f is locally Lipschitz continuous and given $x^0 \in \mathbb{R}^n$, f is prox-regular at each point in the compact set $\text{lev}_{x^0} f$. For all $x \in \text{lev}_{x^0} f$, there exist $\epsilon(x)$ and $a(x)$ such that f is prox-regular at x with respect to $\epsilon(x)$ and $a(x)$. By Proposition 1, f is para-convex on $B(x, \epsilon(x))$ with respect to $a(x)$ for all $x \in \text{lev}_{x^0} f$; i.e. the function $\tilde{g}(y; a(x)) := f(y) + \frac{a(x)}{2}\|y\|^2$ is convex on $B(x, \epsilon(x))$ for each $x \in \text{lev}_{x^0} f$. We see $\{\text{int } B(x, \epsilon(x)) | x \in \text{lev}_{x^0} f\}$ is an open cover of $\text{lev}_{x^0} f$ and it has a finite subcover $\{\text{int } B(x_i, \epsilon(x_i)) | i = 1, \cdots, m\}$ corresponding to some $a(x_i)$, $i = 1, \cdots, m$. Define $\bar{a} := \max\{a(x_i) | i = 1, \cdots, m\}$. Then the function $\tilde{g}(y; \bar{a}) = \tilde{g}(y; a(x_i)) + \frac{\bar{a}-a(x_i)}{2}\|y\|^2$ for all $i = 1, \cdots, m$ is also convex on each $B(x_i, \epsilon(x_i))$. Consequently, $g(y; x, \bar{a}) = \tilde{g}(y; \bar{a}) - \bar{a}\langle y, x\rangle + \frac{\bar{a}}{2}\|x\|^2$ is convex on each $B(x_i, \epsilon(x_i))$ and so is $g(y; x, a) = g(y; x, \bar{a}) + \frac{a-\bar{a}}{2}\|y - x\|^2$ for all $a \geq \bar{a}$. $\square$

The following theorem shows that there exists a threshold for the value a such that the function $g(y)$ is the restriction to $\text{lev}_{x^0} f$ of a convex function. See the Appendix for the proof of this theorem.

Theorem 2. *Suppose f is prox-regular and locally Lipschitz on a bounded level set $\text{lev}_{x^0} f$ with $\text{int }\text{lev}_{x^0} f \neq \emptyset$. Let $g(y; x, a)$ be defined in (17) with $a \geq 0$. There exists a number $a^{th} \geq 0$ such that for all $a \geq a^{th}$, $g(y; x, a)$ is the restriction to $\text{lev}_{x^0} f$ of a globally convex function $H(y; x, a)$ satisfying $g(y; x, a) \geq H(y; x, a)$ for all $y \in \mathbb{R}^n$ and $x \in \text{lev}_{x^0} f$.*

Theorem 2 essentially shows that g can be described as a restriction of a convex function. We will in the future refer to this as the 'restriction property'.

4.2 On-The-Fly Convexification

The threshold value a^{th} is hard to find. Our goal in this section (see also Section 4.4) is to find a lower bound for the parameter a such that $g(y; \bar{x}, a)$ is a restriction of a convex function locally within $\text{lev}_{x^0} f$. We first introduce the convexification technique which first appeared in [8]. Suppose we are at the current iteration point, i.e. the current best candidate for a stationary point of f. We denote this point $\bar{x}$ in general and by x^k when it is necessary to indicate it is in the k-th iteration. A necessary condition for this is that all the cutting planes generated at the points in the subset should be below the graph of g, since a convex function is essentially represented by the point-wise supremum of cutting-plane functions.

Denote $\partial g(y; x, a)$ as the subdifferential of function g with respect to variable y. It follows from the calculus of subdifferential and Assumption 1 that

$$\partial g(y; x, a) = \partial f(y) + a(y - x) \text{ and } \partial g(x; x, a) = \partial f(x) \tag{19}$$

for all y and x in the set $\text{lev}_{x^0} f$. For any $s \in \partial f(y)$ and $y \in \text{lev}_{x^0} f$, clearly we have $s + a(y - x) \in \partial g(y; x, a)$. Consequently, the cutting-plane function of g at the point y_i can be written as

$$h(w, \boldsymbol{y}_i) := h(w; x, a, y_i) : w \mapsto f(y_i) + \frac{a}{2} \|y_i - x\|^2 + \langle s_i + a(y_i - x), w - y_i \rangle, \tag{20}$$

where $s_i \in \partial f(y_i)$. According to Theorem 2, under Assumption 1, if $a \geq a^{th}$, g is a restriction to $\text{lev}_{x^0} f$ of a convex function H minorizing g. Thus a cutting plane of g generated at an auxiliary point $y_i \in \text{int }\text{lev}_{x^0} f$ is the same cutting plane of H generated at y_i; additionally, as H is a convex function minorizing g, this cutting plane is not only below the graph of H but also below that of g. In summary we have

$$g(y; x, a) - h(y; x, a, y_i) \geq 0, \ \forall \ y \in \mathbb{R}^n, \ x \in \text{lev}_{x^0} f, \ y_i \in \text{int }\text{lev}_{x^0} f, \ a \geq a^{th}. \tag{21}$$

We provide a localized convexification process by selecting a collection of points around the current iteration point and verifying the necessary condition for convexity. We let 'a' be variable and I be some index set, and set

$$g(y_j; \bar{x}, a) - h(y_j; \bar{x}, a, y_i) \geq 0, \ \forall \ i, j \in I; \tag{22}$$

to deduce the necessary condition for a: $a \geq \tilde{a}^{\min}$, where

$$\tilde{a}^{\min} := \max_{i,j \in I} \{ -\frac{f(y_i) - f(y_j) - \langle s_j, y_i - y_j \rangle}{\frac{1}{2} \|y_i - y_j\|^2} \}. \tag{23}$$

This value can be negative, so we set $a^{\min} := \max\{\tilde{a}^{\min}, 0\}$. Consequently, for any $a \geq a^{\min}$, (22) holds true.

Note that $a^{\min}$ is dependent on the points indexed in I. Consequently, each time a new auxiliary point y_i is obtained $a^{\min}$ needs to be updated. We also note that, $a^{\min}$, the local lower bound for the convexification parameter, determined by (22) where $\bar{x} \in \text{lev}_{x^0} f$, $y_i \in \text{int }\text{lev}_{x^0} f$, $\forall i \in I$, is not greater than a^{th} that satisfies (21).

4.3 The Model Problem and Model Reduction

The cutting-planes model of $g(y; x, a)$ is defined by

$$m(w) := m(w; x, a, I) : w \mapsto \max_{i \in I} \{ h(w; x, a, y_i) \}, \tag{24}$$

where I is the index set of auxiliary points y_i where cutting planes of function g are generated. Our algorithmic model is defined by (24) and in the remainder of this paper, m refers to the model defined in (24) unless otherwise stated. Suppose we are at $\bar{x}$. To proceed in finding a new candidate, we intend to obtain descent in f by minimizing the cutting-planes model of $g(x; \bar{x}, a)$ over a trust region, i.e. we solve the linear subproblem

$$\min_{x \in \mathbb{R}^n} \quad m(x; \bar{x}, a, I) \quad \text{subject to} \quad \|x - \bar{x}\|_\infty \leq \Delta, \tag{25}$$

which is equivalent to the following problem

$$\min_{(x,z) \in \mathbb{R}^{n+1}} \quad z \tag{26a}$$

$$\text{subject to} \quad f(y_i) + \frac{a}{2}\|y_i - \bar{x}\|^2 + \langle s_i + a(y_i - \bar{x}), x - y_i \rangle \leq z, \ i \in I, \tag{26b}$$

$$\|x - \bar{x}\|_\infty \leq \Delta. \tag{26c}$$

We would like to inspect the reduction of the model function m after we have obtained a new trial point via the linear programming problem (26). Denote a general optimal solution of problem (26) by (x^*, z^*) and by (x^{kl}, z^{kl}) when it is necessary to indicate it is in the l-th minor iteration in the k-th major iteration. In the following lemma we derive the explicit expression of the reduction of the model from $\bar{x}$ to x^*. We will use the linearization errors of $g(\cdot; \bar{x}, a)$ at $\bar{x}$:

$$E_i := g(\bar{x}; \bar{x}, a) - h(\bar{x}; \bar{x}, a, y_i) \tag{27}$$

$$= f(\bar{x}) - \left[f(y_i) + \frac{a}{2}\|y_i - \bar{x}\|^2 + \langle s_i + a(y_i - \bar{x}), \bar{x} - y_i \rangle \right], \ \forall \ i \in I. \tag{28}$$

Lemma 2. *Consider the linear problem* (26). *Let* $\bar{i} \in I$ *be such that* $\bar{x} = y_{\bar{i}}$, $m(x) := m(x; \bar{x}, a, I)$ *be defined as in* (24), (x^*, z^*) *be an optimal solution of* (26) *and suppose 'a' is such that* $E_i \geq 0, \ \forall \ i \in I$.
(i) The following holds true

$$m(\bar{x}) - m(x^*) = f(\bar{x}) - z^* = \begin{cases} \sum_{i \in \bar{I}} \lambda_i E_i, & \text{if } \|x^* - \bar{x}\|_\infty < \Delta; \\ \sum_{i \in \bar{I}} \lambda_i E_i + \Delta\|\sum_{i \in \bar{I}} \lambda_i[s_i + a(y_i - \bar{x})]\|_1, & \text{if } \|x^* - \bar{x}\|_\infty = \Delta, \end{cases} \tag{29}$$

where $\bar{I}$ *is the index set for active constraints in* (26b) *at* (x^*, z^*), *and* λ_i *for* $i \in \bar{I}$ *are the corresponding Lagrangian multipliers of* (26b).
(ii) Let C *be any set satisfying* $\bar{x} \in \text{int } C$ *and* $y_i \in \text{int } C$ *for all* $i \in \bar{I}$, *if additionally 'a' is such that* $g(y; x, a)$ *is a restriction to* C *of a globally convex function* $H(y; x, a)$ *satisfying* $g(y; x, a) \geq H(y; x, a)$ *for all* $y \in \mathbb{R}^n$, $x \in \mathbb{R}^n$. *Then*

$$\sum_{i \in \bar{I}} \lambda_i[s_i + a(y_i - \bar{x})] \in \partial_{\tilde{\epsilon}} g(\bar{x}; \bar{x}, a), \ \text{where } \tilde{\epsilon} = \sum_{i \in \bar{I}} \lambda_i E_i; \tag{30}$$

and furthermore, if $0 \in \partial_0 g(\bar{x}; \bar{x}, a)$, *then* $\bar{x}$ *is a global minimizer of* $g(y; \bar{x}, a)$.

Proof. (i) By definition $m(\bar{x}) = \max_{i \in I}\{h(\bar{x}; \bar{x}, a, y_i)\} \geq h(\bar{x}; \bar{x}, a, y_{\bar{i}})$ and since $\bar{x} = y_{\bar{i}}$, we have $h(\bar{x}; \bar{x}, a, y_{\bar{i}}) = f(y_{\bar{i}}) = f(\bar{x})$ by the definition of function h in (20). As 'a' is such that $E_i \geq 0$ for all $i \in I$, we get $\max_{i \in I}\{h(\bar{x}; \bar{x}, a, y_i)\} \leq g(\bar{x}; \bar{x}, a) = f(\bar{x})$ via the definition of E_i in (27). Consequently, $m(\bar{x}) = f(\bar{x})$. Problem (26) is equivalent to (25) in the sense that the optimal solution (x^*, z^*) satisfies $m(x^*) = z^* = \min\{m(x) \big| \|x - \bar{x}\|_\infty \leq \Delta\}$. Therefore $m(\bar{x}) - m(x^*) = f(\bar{x}) - z^*$. The $\bar{I}$ in (29) is defined by

$$\bar{I} := \left\{ i \in I \big| f(y_i) + \frac{a}{2}\|y_i - \bar{x}\|^2 + \langle s_i + a(y_i - \bar{x}), x^* - y_i \rangle = z^* \right\}. \tag{31}$$

$\bar{I}$ cannot be empty because (x^*, z^*) has to be on some cutting plane. If $\|x^* - \bar{x}\|_\infty = \Delta$, then one of the sets

$$I_R := \left\{ i \in \{1, \cdots, n\} \big| x_i^* = \bar{x}_i + \Delta \right\}, \text{ and } I_L := \left\{ i \in \{1, \cdots, n\} \big| x_i^* = \bar{x}_i - \Delta \right\} \tag{32}$$

will be nonempty. As (x^*, z^*) is the optimal solution of linear problem (26), it satisfies the Karush-Kuhn-Tucker (KKT) conditions; that is, there exist multipliers $\lambda_i \geq 0$, $i \in \bar{I}$, $u_i \geq 0$, $i \in I_R$, and $w_i \geq 0$, $i \in I_L$ such that

$$\sum_{i \in \bar{I}} \lambda_i[s_i + a(y_i - \bar{x})] + \sum_{i \in I_R} u_i e_i - \sum_{i \in I_L} w_i e_i = 0, \tag{33}$$

$$1 - \sum_{i \in \bar{I}} \lambda_i = 0, \tag{34}$$

where $e_i, \ i = 1, \cdots, n$ are vectors in $\mathbb{R}^n$ with the i-th component being 1 and the others 0. If $\|x^* - \bar{x}\|_\infty < \Delta$, then (33) simply reduces to

$$\sum_{i \in \bar{I}} \lambda_i [s_i + a(y_i - \bar{x})] = 0. \tag{35}$$

We have

$$\begin{aligned}
f(\bar{x}) - z^* &= \sum_{i \in \bar{I}} \lambda_i \left[f(\bar{x}) - z^* \right] \text{ (by (34))} \\
&= \sum_{i \in \bar{I}} \lambda_i \left\{ f(\bar{x}) - \left[f(y_i) + \frac{a}{2} \|y_i - \bar{x}\|^2 + \langle s_i + a(y_i - \bar{x}), x^* - y_i \rangle \right] \right\} \text{ (by (31))} \\
&= \sum_{i \in \bar{I}} \lambda_i \left\{ f(\bar{x}) - \left[f(y_i) + \frac{a}{2} \|y_i - \bar{x}\|^2 + \langle s_i + a(y_i - \bar{x}), \bar{x} - y_i \rangle \right] \right\} \\
&\qquad\qquad\qquad\qquad\qquad - \sum_{i \in \bar{I}} \lambda_i \langle s_i + a(y_i - \bar{x}), x^* - \bar{x} \rangle. \tag{36}
\end{aligned}$$

Consider the first case when $\|x^* - \bar{x}\|_\infty < \Delta$. Then by (28), (35) and (36) we have

$$f(\bar{x}) - z^* = \sum_{i \in \bar{I}} \lambda_i E_i.$$

If $\|x^* - \bar{x}\|_\infty = \Delta$, then by (28), (33) and (36) we have

$$\begin{aligned}
f(\bar{x}) - z^* &= \sum_{i \in \bar{I}} \lambda_i E_i + \left\langle \sum_{i \in I_R} u_i e_i - \sum_{i \in I_L} w_i e_i, x^* - \bar{x} \right\rangle \\
&= \sum_{i \in \bar{I}} \lambda_i E_i + \Delta \left(\sum_{i \in I_R} u_i + \sum_{i \in I_L} w_i \right) \text{ (by definition of } I_R \text{ and } I_L\text{)}. \tag{37}
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\| \sum_{i \in \bar{I}} \lambda_i [s_i + a(y_i - \bar{x})] \|_1 &= \| \sum_{i \in I_R} u_i e_i - \sum_{i \in I_L} w_i e_i \|_1 \text{ (by (33))} \\
&= \left(\sum_{i \in I_R} u_i + \sum_{i \in I_L} w_i \right) \text{ (by definition of } I_R \text{ and } I_L\text{)}. \tag{38}
\end{aligned}$$

Combining (37) and (38) we get

$$f(\bar{x}) - z^* = \sum_{i \in \bar{I}} \lambda_i E_i + \Delta \| \sum_{i \in \bar{I}} \lambda_i [s_i + a(y_i - \bar{x})] \|_1.$$

(ii) By the restriction property in Theorem 2 we have $g(y; \bar{x}, a) = H(y; \bar{x}, a)$ for all $y \in C$, and $\partial g(y; \bar{x}, a) = \partial H(y; \bar{x}, a)$ for all $y \in \text{int } C$. As H is globally convex, $\partial_\epsilon g(y; \bar{x}, a)$ can be defined by $\partial_\epsilon g(y; \bar{x}, a) := \partial_\epsilon H(y; \bar{x}, a)$ for any $y \in \text{int } C$. For all $i \in \bar{I}$, since $y_i \in \text{int } C$, we have $s_i + a(y_i - \bar{x}) \in \partial H(y_i; \bar{x}, a)$. Thus we have that

$$\begin{aligned}
g(z; \bar{x}, a) &\geq H(z; \bar{x}, a) \text{ for all } z \in \mathbb{R}^n \\
&\geq H(y_i; \bar{x}, a) + \langle s_i + a(y_i - \bar{x}), z - y_i \rangle \text{ for all } i \in \bar{I} \text{ and } z \in \mathbb{R}^n \\
&\quad \text{(by definition of convex subgradient)} \\
&= g(y_i; \bar{x}, a) + \langle s_i + a(y_i - \bar{x}), z - y_i \rangle \text{ (by the restriction property)} \\
&= g(\bar{x}; \bar{x}, a) + \langle s_i + a(y_i - \bar{x}), z - \bar{x} \rangle - E_i \text{ (by definition of } E_i\text{)}. \tag{39}
\end{aligned}$$

The convex combination of (39) with λ_i satisfying (34) yields

$$g(z; \bar{x}, a) \geq g(\bar{x}; \bar{x}, a) + \left\langle \sum_{i \in \bar{I}} \lambda_i [s_i + a(y_i - \bar{x})], z - \bar{x} \right\rangle - \sum_{i \in \bar{I}} \lambda_i E_i, \text{ for all } z \in \mathbb{R}^n. \tag{40}$$

By the definition of $\epsilon-$subdifferential in convex analysis, (40) verifies (30).

If $0 \in \partial_0 g(\bar{x}; \bar{x}, a)$, then $0 \in \partial H(\bar{x}; \bar{x}, a)$ since $\bar{x} \in \text{int } C$ and $\partial H(y; \bar{x}, a) = \partial g(y; \bar{x}, a)$ for all $y \in \text{int } C$. Because $H(y; \bar{x}, a)$ is a convex function, its stationary points are also global minimizers. Consequently, for any $z \in \mathbb{R}^n$, we have $g(\bar{x}; \bar{x}, a) = H(\bar{x}; \bar{x}, a) \leq H(z; \bar{x}, a) \leq g(z; \bar{x}, a)$. We note that another simple way to see this conclusion is that $0 \in \partial g(\bar{x}; \bar{x}, a) = \partial f(\bar{x})$ which is equivalent to $\bar{x} \in P_a(\bar{x})$. $\qquad \square$

Remark. (i) The conclusions in Lemma 2 are very similar to those of the classical bundle methods. The convex combination of subgradients, $\sum_{i \in \bar{I}} \lambda_i s_i$ (or $\sum_{i \in \bar{I}} \lambda_i [s_i + a(y_i - \bar{x})]$) is involved in both situations and is sometimes termed an *aggregate subgradient*. The *subgradient aggregation* technique was developed in [13] where aggregate subgradients together with the convex combinations of linearization errors are used to represent additional virtual cutting planes in the model. Subgradient aggregation technique has many applications including preventing unbounded storage caused by too many cutting planes.

(ii) It is not difficult to see that the model reduction is always nonnegative provided that $E_i \geq 0$ for all $i \in I$. From the definition of E_i in (27), this can be guaranteed by choosing $a \geq a^{\min}$ which satisfies (22).

(iii) The expression of model reduction in (29) can also be stated as

$$m(\bar{x}) - m(x^*) = \sum_{i \in \bar{I}} \lambda_i E_i + \Delta \| \sum_{i \in \bar{I}} \lambda_i [s_i + a(y_i - \bar{x})] \|_1 \tag{41}$$

with $\sum_{i \in \bar{I}} \lambda_i [s_i + a(y_i - \bar{x})] = 0$ if $\|x^* - \bar{x}\|_\infty < \Delta$.

Lemma 2 implies that the model reduction can somehow help us to determine whether $\bar{x}$ is a good estimate of a stationary point. If f is convex, generally speaking, a good estimate $\bar{x}$ of a minimizer of f should satisfy that both $\min_{g \in \partial f(\bar{x})} \|g\|$ and $f(\bar{x}) - f(p)$ are very small, where p is a minimizer of f. In the convex case $\frac{f(\bar{x}) - f(p)}{\|\bar{x} - p\|_\infty}$ can be a lower bound for $\min_{g \in \partial f(\bar{x})} \|g\|$. Motivated by this, in the next lemma we try to relate the model reduction with the above two approximate measures of a good estimate of a stationary point in the nonconvex case through the restricted convexity. We will use the following set which essentially defines the model $m(x)$.

$$F := \{y_i \mid i \in \hat{I}\}, \text{ where } \hat{I} := \{i \in I \mid \exists \, x \in \mathbb{R}^n \text{ such that } m(x) = h(x; \bar{x}, a, y_i)\}. \tag{42}$$

Lemma 3. *Given a prox-center $\bar{x} \in \text{lev}_{x^0} f$, a trust region radius Δ and an index set I containing some $\bar{i}$ such that $\bar{x} = y_{\bar{i}}$. Let $m(x)$ defined in (24) be the objective function of problem (25), 'a' be such that $E_i \geq 0, \, \forall \, i \in I$, x^* be the first component of an optimal solution of problem (26), and F be defined in (42).*

If 'a' is such that $g(x; \bar{x}, a)$ is a restriction to set F of $H(x)$, where $H(x)$ is convex on $\mathbb{R}^n$ and satisfies $H(x) \leq g(x)$ for all $x \in \mathbb{R}^n$. Then
(a) $m(x)$ is a cutting-plane model of $H(x)$ and satisfies $m(x) \leq H(x), \, \forall \, x \in \mathbb{R}^n$;
(b) if $p \in P_a(\bar{x})$ and $\bar{x} \notin P_a(\bar{x})$, then

$$m(\bar{x}) - m(x^*) \geq [f(\bar{x}) - e_a(\bar{x})] \min\{\frac{\Delta}{\|\bar{x} - p\|_\infty}, 1\} \geq 0. \tag{43}$$

Proof. (a) We see that $\hat{I}$ indexes all the cutting planes that sufficiently define $m(x)$. The model $m(x)$ is essentially the pointwise maximum of cutting planes of $g(x)$ generated at bundle points where the value of g and H coincide, and hence $m(x)$ is also a cutting-plane model of $H(x)$. Since $H(x)$ is convex, by proposition 2 it is a lower approximation of H. This finishes the proof of (a).

(b) The proof of (43) can be divided into two parts based on the possible positions of p. First, suppose p is located in the trust region, i.e. $\|\bar{x} - p\|_\infty \leq \Delta$, which yields $\min\{\frac{\Delta}{\|\bar{x} - p\|_\infty}, 1\} = 1$. To show (43) it suffices to show $m(\bar{x}) - m(x^*) \geq f(\bar{x}) - e_a(\bar{x})$. By Lemma 2 (i), $m(\bar{x}) = f(\bar{x})$, and hence we only need to show $m(x^*) \leq e_a(\bar{x})$. From the optimality of x^*, conclusion (a), the fact that $H(x) \leq g(x)$ and 18, we have

$$m(x^*) \leq m(p) \leq H(p) \leq g(p) = e_a(\bar{x}).$$

Second, suppose p is outside the trust region, i.e. $\|\bar{x} - p\|_\infty > \Delta$, which yields $\min\{\frac{\Delta}{\|\bar{x} - p\|_\infty}, 1\} = \frac{\Delta}{\|\bar{x} - p\|_\infty}$. To show (43) it suffices to show

$$m(\bar{x}) - m(x^*) \geq \frac{f(\bar{x}) - e_a(\bar{x})}{\|\bar{x} - p\|_\infty} \Delta. \tag{44}$$

We consider the point $x^c = \bar{x} + \frac{\Delta}{\|\bar{x}-p\|_\infty}(p-\bar{x})$, the intersection point of trust region and the line segment $[\bar{x}, p]$. By the optimality of x^*, the result (a), the convexity of H, the fact that $H(x) \leq g(x)$, (18) and the fact that $g(\bar{x}) = f(\bar{x})$, we have

$$
\begin{aligned}
m(x^*) \leq m(x^c) \leq H(x^c) &\leq \frac{\Delta}{\|\bar{x}-p\|_\infty}H(p) + (1 - \frac{\Delta}{\|\bar{x}-p\|_\infty})H(\bar{x}) \\
&\leq \frac{\Delta}{\|\bar{x}-p\|_\infty}g(p) + (1 - \frac{\Delta}{\|\bar{x}-p\|_\infty})g(\bar{x}) \\
&= \frac{\Delta}{\|\bar{x}-p\|_\infty}e_a(\bar{x}) + (1 - \frac{\Delta}{\|\bar{x}-p\|_\infty})f(\bar{x}).
\end{aligned}
$$

Then (44) can be verified using $f(\bar{x}) = m(\bar{x})$. $\qquad\square$

4.4 Update of the Model

At the end of a certain iteration (either major or minor), we need to update the model and prepare the data for the new LP in next iteration. The update of the model is supposed to improve the model. The update includes adding new cutting planes and deleting old cutting planes, i.e. adding and removing points from the set $\Omega := \{y_i \mid i \in I\}$. We also take into account the update of convexification parameter a and $a^{\min}$ when considering updating the model, as a is also part of the model.

In our method, we always add one cutting plane at the end of each iteration. Specifically, at the end of a major iteration, we obtain x^{k+1} as our new prox-center. A cutting plane will be needed to generate at this point, and hence we add x^{k+1} to Ω so that there exists a $\bar{i} \in I$ such that $y_{\bar{i}} = x^{k+1}$. At the end of a minor iteration, we obtain x^{kl} which did not yield sufficient reduction of the objective function. A cutting plane is supposed to be generated at this point to improve the quality of the model. However x^{kl} could be very bad in the sense that $f(x^{kl})$ is too far away from $f(x^k)$ or even bigger than $f(x^0)$. In this case we backtrack along the direction $x^{kl} - x^k$ until we find a point whose funciton value is less than some upper bound f_u^k. This is a finite process provided that $f_u^k > f^k$, as we will prove in Lemma 5.

After backtrack, we add the point found into Ω, and consequently, all our new bundle points, i.e. those that are generated in iteration (k, l) for some l, will be in $\text{lev}_{f_u^k}f$. However the old bundle points, i.e. those generated in $(k-j, l')$ for some j and l', can still be outside $\text{lev}_{f_u^k}f$. At the end of iteration (k, l_k), we remove the old bundle points whose function values are greater than f_u^k. Finally, before we enter iteration $k+1$ we move the upper bound closer to $f(x^{k+1})$ by setting $f_u^{k+1} \leftarrow \alpha_3 f^k + (1 - \alpha_3)f_u^k$ with some $\alpha_3 \in (0, 1)$.

We see the bundle point set Ω is updated dynamically so that at any iteration (k, l), $\Omega \subseteq \text{lev}_{f_u^{k-2}}f$. This setting is related to the convexification process. In Lemma 3 the value of parameter a such that g is a restriction of a convex function depends on the set $F \subseteq \Omega$. The following lemma states the existence of such value.

Lemma 4. *Suppose that Assumption 1 holds. Given a prox-center $\bar{x} \in \text{lev}_{x^0}f$ and some $f_u \in (f(\bar{x}), f(x^0)]$, consider the model $m(x)$ defined in (24) with bundle points $y_i \in \Omega$ satisfying $y_i \in \text{lev}_{f_u}f$ for all $i \in I$. Let D be a compact set such that $D \supseteq F$ and $\text{int } D \neq \emptyset$. There exists a threshold $a^{th}(\bar{x}, I) \geq 0$ such that for all $a \geq a^{th}(\bar{x}, I)$, $g(y; x, a)$ is the restriction to D of a globally convex function $H(y; x, a)$ satisfying $g(y; x, a) \geq H(y; x, a)$ for all $y \in \mathbb{R}^n$ and $x \in D$.*

Proof. This lemma is an extension of Theorem 2. Since $y_i \in \text{lev}_{f_u}f$ for all $i \in I$ and $f_u \in (f(\bar{x}), f(x^0)]$ we have by (42) that $F \subseteq \text{lev}_{x^0}f$. As F is a finite set and D is the smallest compact set containing F we have $D \subseteq \text{lev}_{x^0}f$. From Corollary 1 and the poof of Theorem 2 in the appendix, we can see that the same conclusion holds true when we replace $\text{lev}_{x^0}f$ by any of its compact subset that has nonempty interior. In our case each $(\bar{x}, I)$ corresponds to a set $D \subseteq \text{lev}_{x^0}f$ and for each D there exists a threshold $a^{th}(\bar{x}, I)$ satisfying the corresponding conditions. $\qquad\square$

From Lemma 4 we see that the condition for 'a' in Lemma 3 can be satisfied if we take $a \geq a^{th}(\bar{x}, I)$.

4.5 The LPBNC Algorithm

For the trust region update we follow the procedure 1. In our algorithm in order to distinguish the prox-center we differentiate major iterations and minor iterations. When we set a new prox-center x^{k+1} it is also the last minor iteration point denoted by x^{kl_k}. As we have an infinite sequence of iterations, the following two situations

can happen. First, there are infinite number of major iterations with each loop of minor iteration to be finite. The sequence can be described as follows:

$$x^0, \ x^{00}, \ x^{01}, \cdots, \ x^1(= x^{0l_0}), \ x^{10}, \ x^{11}, \ \cdots, \ x^k, \ x^{k0}, \ x^{k1}, \ \cdots, \ x^{k+1}(= x^{kl_k}), \ \cdots\cdots. \tag{45a}$$

Second, there are finite number of major iterations with the last major iteration containing infinite minor iterations, which can be described as:

$$x^0, \ x^{00}, \ x^{01}, \cdots, \ x^1(= x^{0l_0}), \ x^{10}, \ x^{11}, \ \cdots, \ x^k, \ x^{k0}, \ x^{k1}, \ \cdots, \ x^{kl}, \ x^{k(l+1)}, \ \cdots\cdots. \tag{45b}$$

A set of similar notations goes for the sequences of parameters (a, Δ). Starting from (a_0^0, Δ_0^0), each pair (a_l^k, Δ_l^k) is used to produce x^{kl}; and if $l = l_k$ we say x^{kl_k} was produced by $(a_{l_k}^k, \Delta_{l_k}^k)$. To alleviate notation we drop the subscripts of (a_l^k, Δ_l^k) in the Algorithm 3 below and also in later analysis we define $m_l^k(x) := m(x; x^k, a_l^k, I(k, l))$ for all k and l and note the value of $m_l^k(x)$ is dependent on x^k, a_l^k and $I(k, l)$.

Algorithm 3: LPBNC

Data: Final accuracy tolerance ϵ_{tol}, maximum trust region radius $\Delta_{\max}$, initial trust region $\Delta_0^0 \in (0, \Delta_{\max})$, initial point x^0, trust region parameters $0 < \eta_1 < \eta_3 < 1$ and $0 < \alpha_1 < 1 < \alpha_2$, backtrack parameter $\beta \in (0,1)$, parameter $\sigma \in \mathbb{R}_{\geq 1}$ and increasing parameter for convexification parameter $\gamma \in [2, 10]$.

1 **Initialization** major and minor iteration counter $(k,l) \leftarrow (0,0)$, initial convexification parameter $a \leftarrow 0$, $a^{\min} \leftarrow 0$, add x^k into Ω, generate a cutting plane of $g(y; x^k, a)$ at x^k, prepare the information for the first LP, $f_u^0 \leftarrow f(x^0)$;

2 solve the linear programming subproblem (26) with $\bar{x}$, I replaced by x^k, $I(k,l)$ and obtain an optimal solution (x^{kl}, z^{kl});

3 **if** $f(x^k) - z^{kl} \leq (1 + |f(x^k)|)\epsilon_{\text{tol}}$ **then**

4 $\quad$ STOP;

5 **end**

6 **if** $\rho_l^k = \frac{f(x^k) - f(x^{kl})}{f(x^k) - z^{kl}} \geq \eta_1$ **then**

7 $\quad$ $serious \leftarrow 1$, $x^{k+1} \leftarrow x^{kl}$, $l_k \leftarrow l$;

$\quad$ /* update trust region radius for major iteration $\qquad\qquad\qquad\qquad\qquad\qquad$ */

8 $\quad$ **if** $\rho_l^k > \eta_3$ **and** $\|x^{kl} - x^k\|_\infty > 0.9\Delta$ **then**

9 $\quad\quad$ $\Delta \leftarrow \min\{\alpha_2\Delta, \Delta_{\max}\}$

10 $\quad$ **end**

11 $\quad$ for all $w \in \Omega$ if $w \notin \text{lev}_{f_u^k} f$, delete w from Ω and update the index set $I(k,l)$ by deleting the index whose corresponding cutting plane is generated at w ;

12 $\quad$ $f_u^{k+1} \leftarrow \alpha_3 f(x^{k+1}) + (1 - \alpha_3) f_u^k$

13 **else**

14 $\quad$ $serious \leftarrow 0$;

$\quad$ /* update trust region radius for minor iteration $\qquad\qquad\qquad\qquad\qquad\qquad$ */

15 $\quad$ **if** $\rho_l^k < -\frac{1}{\min\{1,\Delta\}}$ **then**

16 $\quad\quad$ $\Delta \leftarrow \alpha_1\Delta$

17 $\quad$ **end**

18 $\quad$ **if** $k > 0$ **and** $f(x^{kl}) > f_u^k$ **then** // backtrack

19 $\quad\quad$ $\boldsymbol{d} \leftarrow x^{kl} - x^k; j \leftarrow 1$;

20 $\quad\quad$ **while** $f(x^k + \beta^j \boldsymbol{d}) > f_u^k$ **do**

21 $\quad\quad\quad$ $j \leftarrow j + 1$

22 $\quad\quad$ **end**

23 $\quad\quad$ $\bar{\boldsymbol{y}} \leftarrow x^k + \beta^j \boldsymbol{d}$

24 $\quad$ **end**

25 **end**

26 add x^{k+1} in major iteration, or add x^{kl}, or $\bar{y}$ if backtrack was performed, as a new bundle point y_i into Ω, update $a^{\min}$ according to its definition in (23);

27 **if** $a < a^{\min}$ **then**

28 $\quad$ $a \leftarrow \max\{a^{\min}, \gamma a\}$

29 **else if** $a^{\min} > 0$ **and** $a \geq \sigma a^{\min}$ **then**

30 $\quad$ $a \leftarrow (a + a^{\min})/2$

31 **end**

32 **if** $serious$ **then**

33 $\quad$ generate a cutting plane of $g(y; x^{k+1}, a)$ at x^{k+1}, and add the cutting-plane function to the model. (The model becomes m_0^{k+1} and the cutting plane index set becomes $I(k+1, l)$.) $k \leftarrow k + 1$, $l \leftarrow 0$;

34 **else**

35 $\quad$ generate a cutting plane of $g(y; x^k, a)$ at the new bundle point, and add the cutting-plane function to the model. (The model becomes m_{l+1}^k and the cutting plane index set becomes $I(k, l+1)$.)

36 $\quad$ $l \leftarrow l + 1$

37 **end**

38 update the coefficient matrix in LP subproblem (26);

39 continue to next iteration by going to line 2

Remark. 1. The value of a used in the LP subproblem (26) can always guarantee $E_i \geq 0$ for all $i \in I(k, l)$.

2. We set x^{k+1} as x^{kl} if $\rho_l^k \geq \eta_1$. This implies $f(x^{k+1}) < f(x^k)$ and thus $x^k \in \text{lev}_{x^0} f$ for all k.

3. At the beginning of iteration k, a cutting plane of $g(\cdot; x^k, a)$ is generated at x^k. From the update of f_u^k we see that $f(x^k) \leq f_u^k$ for all k. At the end of iteration k, x^k will not be deleted from Ω and consequently x^k is always indexed in $I(k, l)$ for all $0 \leq l \leq l_k$.

4.6   LPBNC is well-defined

The following lemma shows that if x^{kl} is not located in $\text{lev}_{f_u^k} f$ then after finite backtrack along the direction $x^{kl} - x^k$, we can reach an auxiliary point in $\text{lev}_{f_u^k} f$.

Lemma 5. *If at iteration (k, l) we have $f(x^{kl}) > f_u^k$ and $k > 0$, then there exists an integer j' such that $\bar{y} = x^k + \beta^{j'} \boldsymbol{d}$ and $\bar{y} \in \text{lev}_{f_u^k} f$ with $\boldsymbol{d} = x^{kl} - x^k$.*

Proof. From the algorithm we see $f_u^0 = f(x^0) > f(x^k)$ and $f_u^k = \alpha_3 f(x^k) + (1 - \alpha_3) f_u^{k-1}$ if $k > 0$. Hence, $f_u^k - f(x^k) > 0$ for all $k > 0$. Suppose for contradiction that $f(x^k + \beta^j \boldsymbol{d}) > f_u^k$ for all integer $j > 0$. Hence, $0 < f_u^k - f(x^k) < f(x^k + \beta^j \boldsymbol{d}) - f(x^k) \leq \beta^j \|\boldsymbol{d}\|_\infty \bar{L}$ where $\bar{L}$ is the Lipschitz constant of f. Therefore $\beta^j > \frac{f_u^k - f(x^k)}{\|\boldsymbol{d}\|_\infty \bar{L}} := c_1$. Since $\beta \in (0, 1)$ clearly this cannot be true for all $j > 0$; a contradiction. $\square$

We want to show that the algorithm LPBNC is well defined by showing that the inner loop (loop of minor iterations) can terminate finitely and that if it does not terminate finitely, then we have already found a stationary point of f.

Lemma 6. *Suppose $\rho_l^k < \eta_1$ for some k, l. Then*

$$m_{l+1}^k(x^{k,l+1}) \geq m_l^k(x^{kl})$$

Proof. At the end of iteration k, l, we do not delete any cutting planes but add a new cutting plane to the model. Furthermore, the trust region radius Δ is possibly decreased. Hence the feasible region of linear subproblem (26) for iteration $k, l + 1$ will become smaller. Therefore we have $m_{l+1}^k(x^{k,l+1}) \geq m_l^k(x^{kl})$. $\square$

Define

$$L = \sup\{\|s\|_1 : s \in \partial f(y), \ \|y - x\|_\infty \leq \Delta_{\max}, \ x \in \text{lev}_{x^0} f\} \tag{46}$$

From the Lipschitz continuity and prox-regularity of f we have $L < +\infty$.

The notion of *minor iteration* is similar to the *null step* in bundle methods. The following lemma shows that minor iterations either terminate finitely or generate an infinite sequence with very small model reduction. Furthermore, as we can see from below, the model reduction eventually decreases to 0 if minor iterations do not terminate finitely. If f is convex we can easily see that this will show that the current iteration point is already a global minimizer of f. For the nonconvex case, we will show that during the infinite minor iterations, if the convexification succeeds, i.e. the function g is eventually convex locally around x^k, then x^k is a stationary point of f.

Lemma 7. *Suppose that $\epsilon_{\text{tol}} = 0$ and a_l^k is bounded above by a constant A for all k and l. Let l_1 be any index such that $\rho_{l_1}^k < \eta_1$. Then there is an index $l_2 > l_1$ and a real number $\bar{\eta}_2 \in (\eta_1, 1)$ such that either $\rho_{l_2}^k \geq \eta_1$ or*

$$\frac{f(x^k) - m_{l_2}^k(x^{kl_2})}{f(x^k) - m_{l_1}^k(x^{kl_1})} \leq \bar{\eta}_2. \tag{47}$$

Proof. Suppose for contradiction that there does not exist such index l_2 and real number $\bar{\eta}_2$; that is, there is an infinite sequence of minor iterations and

$$\frac{f(x^k) - m_q^k(x^{kq})}{f(x^k) - m_{l_1}^k(x^{kl_1})} > \bar{\eta}_2, \quad \forall q > l_1, \ \forall \bar{\eta}_2 \in (\eta_1, 1). \tag{48}$$

Since l_1 can be any index such that $\rho_{l_1}^k < \eta_1$, and we do not delete cutting planes in minor iterations, we can assume that q and l are generic indices satisfying $q > l \geq l_1$.

14

To construct a contradiction, write $f(x^k) - m_q^k(x) = [f(x^k) - m_q^k(x^{kl})] + [m_q^k(x^{kl}) - m_q^k(x)]$. Consider the two parts of the right hand side of this equation. First, observing that x^{kl} is the point where a new cutting plane of $g(y; x^k, a_l^k)$ is generated and all the cutting planes in minor iterations are kept in the model, it follows that $g(x^{kl}; x^k, a_q^k) \leq m_q^k(x^{kl})$. Consequently,

$$f(x^k) - m_q^k(x^{kl}) \leq f(x^k) - g(x^{kl}; x^k, a_q^k) = f(x^k) - \left[f(x^{kl}) + \frac{a_q^k}{2} ||x^{kl} - x^k||^2 \right]$$

$$\leq f(x^k) - f(x^{kl}) < \eta_1 \left[f(x^k) - m_l^k(x^{kl}) \right] \quad \text{(becasue } \rho_l^k < \eta_1)$$

$$\leq \eta_1 \left[f(x^k) - m_{l_1}^k(x^{kl_1}) \right] \quad \text{(by Lemma 6)}. \tag{49}$$

Second, the model function $m_q^k(x)$ is convex. Therefore for all $\tilde{s} \in \partial m_q^k(x^{kl})$,

$$m_q^k(x) - m_q^k(x^{kl}) \geq \tilde{s}^T(x - x^{kl}) \quad \forall x. \tag{50}$$

Note $\tilde{s}$ is a subgradient of m_q^k at x^{kl} where a cutting plane of $g(x^{kl}; x^k, a_q^k)$ was generated, and thus $\tilde{s} \in \partial g(x^{kl}; x^k, a_q^k) = \partial f(x^{kl}) + a_q^k(x^{kl} - x^k)$. It follows from (46), (26c) and the boundedness of a_l^k that $||\tilde{s}||_1 \leq L + ||a_q^k(x^{kl} - x^k)||_1 \leq L + An\Delta_{\max}$. Applying this to (50) we have

$$m_q^k(x^{kl}) - m_q^k(x) \leq ||\tilde{s}||_1 ||x - x^{kl}||_\infty \leq (L + An\Delta_{\max})||x - x^{kl}||_\infty \quad \forall x. \tag{51}$$

Summing (49) and (51), there is

$$f(x^k) - m_q^k(x) < \eta_1[f(x^k) - m_{l_1}^k(x^{kl_1})] + (L + An\Delta_{\max})||x - x^{kl}||_\infty \quad \forall x. \tag{52}$$

It then follows from (48) and (52) by taking x as x^{kq} that

$$\bar{\eta}_2[f(x^k) - m_{l_1}^k(x^{kl_1})] < \eta_1[f(x^k) - m_{l_1}^k(x^{kl_1})] + (L + An\Delta_{\max})||x^{kq} - x^{kl}||_\infty \quad \forall q > l_1.$$

Thus

$$||x^{kq} - x^{kl}||_\infty > \frac{\bar{\eta}_2 - \eta_1}{(L + An\Delta_{\max})}[f(x^k) - m_{l_1}^k(x^{kl_1})] := c_2 > 0. \tag{53}$$

However this cannot happen for an infinite number of indices q and l because all minor iteration points x^{kl} such that $l \geq l_1$ are in the neighborhood of x^k, $B(x^k, \Delta_{\max}) = \{x \mid ||x - x^k||_\infty \leq \Delta_{\max}\}$. Hence a contradiction is found and (47) must be true for some $l_2 > l_1$ and $\bar{\eta}_2 > \eta_1$. $\square$

It is worthy to mention that the proof of lemma 6 and 7 does not require the convexity of f. So far we have shown that the minor iterations either terminate finitely or continue infinitely. To demonstrate our algorithm is well defined, we need to show in the latter case that the current major iteration point is a stationary point of f and that a_l^k is indeed bounded above. We show this in the next section.

5 Convergence analysis

Theorem 3. *Let Assumption 1 hold and $\epsilon_{tol} = 0$. Suppose Algorithm 3 terminates at iteration (k, l). If $a_l^k \geq a^{th}(x^k, I(k, l))$ then $x^k \in P_{a_l^k}(x^k)$ and $0 \in \partial f(x^k)$.*

Proof. As the algorithm terminates at x^{kl}, x^{kl} must satisfy the stopping criterion $f(x^k) - z^{kl} \leq (1 + |f(x^k)|)\epsilon_{tol}$. As $\epsilon_{tol} = 0$, we have $f(x^k) - z^{kl} \leq 0$. However by the expression of model reduction in (29), $f(x^k) - z^{kl} \geq 0$. Thus $f(x^k) - z^{kl} = 0$. Suppose for contradiction $x^k \notin P_{a_l^k}(x^k)$. By (29), Lemma 4, and Lemma 3, if $a_l^k \geq a^{th}(x^k, I(k, l))$ and $p \in P_{a_l^k}(x^k)$, then we have

$$0 = f(x^k) - z^{kl} \geq \left[f(x^k) - e_{a_l^k}(x^k) \right] \min\left(\frac{\Delta_l^k}{||x^k - p||_\infty}, 1 \right) \geq 0. \tag{54}$$

From the definition of $P_{a_l^k}(x^k) \ni p$ we have

$$f(x^k) - e_{a_l^k}(x^k) = g(x^k; x^k, a_l^k) - g(p; x^k, a_l^k) > 0, \text{ if } x^k \notin P_{a_l^k}(x^k). \tag{55}$$

From (54) and (55) we see $0 = \min\left\{ \frac{\Delta_l^k}{||x^k - p||_\infty}, 1 \right\}$. Thus $0 = \frac{\Delta_l^k}{||x^k - p||_\infty}$ with $||x^k - p||_\infty \geq \Delta_l^k$. But Δ_l^k cannot be reduced to 0 after finite iterations. Hence we have a contradiction. $\square$

Let ϵ_{tol} be 0. In the following analysis we assume the algorithm does not stop finitely. We see in (45) that two sequences can be generated. Consistent notations should be used for a and $a^{\min}$. For clear understanding, we unify those two cases with $\{a^n\}$ and $\{a_n^{\min}\}$ when it's not necessary to distinguish them. We start our convergence analysis by showing that the convexification parameter a in algorithm LPBNC is bounded.

Lemma 8. $0 \le a^n \le \gamma a^{th}, \ \forall \ n \in \mathbb{N}$.

Proof. The update of a^n happens only in line 28 or line 30 of Algorithm 3. In either case we have $a^n \ge a_n^{\min} \ge 0$ from the definition of $a_n^{\min}$. We increase a in line 28 and decrease a in line 30. To show $a^n \le \gamma a^{th}$ we only need to show $\max\{a_n^{\min}, \gamma a^n\} \le \gamma a^{th}$ if $a^n < a_n^{\min}$. As $\gamma \ge 2$ and $a_n^{\min} \le a^{th}$ for all $n \in \mathbb{N}$, we clearly have $\max\{a_n^{\min}, \gamma a^n\} \le \gamma a_n^{\min} \le \gamma a^{th}$. $\qquad\square$

A consequence of lemma 7 is that if the minor iteration sequence does not terminate finitely then the model reduction will become smaller and smaller and eventually converge to 0. We will show that if there is an infinite sequence of serious steps, the model reduction will converge to 0 too. Denote the index of the last minor iteration as l_k so that $x^{k+1} = x^{kl_k}$.

Lemma 9. *The model reduction of LPBNC converges to 0. Specifically,*
(i) if in iteration k there is an infinite sequence of minor iterations then

$$\lim_{l \to \infty} [m_l^k(x^k) - m_l^k(x^{kl})] = 0; \tag{56}$$

(ii) if the sequence of major iteration points $\{x^k\}$ is infinite then

$$\lim_{k \to \infty} [m_{l_k}^k(x^k) - m_{l_k}^k(x^{k+1})] = 0. \tag{57}$$

Proof. (i) From Lemma 2 we know $m_l^k(x^k) = f(x^k)$ for all (k, l) and the sequence $\{f(x^k) - m_l^k(x^{kl})\}_{l=1}^{\infty}$ is non-negative. From Lemma 6 we see this sequence is also monotonic. Since the sequence is infinite, by Lemma 7 there is an infinite sequence of indices $0 < l_1 < l_2 < \cdots$ such that $0 \le f(x^k) - m_{l_j}^k(x^{kl_j}) \le \bar{\eta}_2[f(x^k) - m_{l_{j-1}}^k(x^{kl_{j-1}})] \le \cdots \le \bar{\eta}_2^{j-1}[f(x^k) - m_{l_1}^k(x^{kl_1})]$ where j can be infinitely large. Consequently, (56) holds true.
(ii) From Remark 39(2) we see the sequence $\{f(x^k)\}_{k=0}^{\infty}$ is monotonic. Under Assumption 1, f is bounded below. Hence $\lim_{k \to \infty} [f(x^k) - f(x^{k+1})] = 0$.
From the definition of $\rho_{l_k}^k$ and Lemma 2, we have $f(x^k) - f(x^{kl_k}) \ge \eta_1[f(x^k) - z^{kl_k}] = (m_{l_k}^k(x^k) - m_{l_k}^k(x^{kl_k}))$. Since both $\{f(x^k) - f(x^{k+1})\}$ and $\{m_{l_k}^k(x^k) - m_{l_k}^k(x^{k+1})\}$ are nonnegative we must have $\lim_{k \to \infty} [m_{l_k}^k(x^k) - m_{l_k}^k(x^{kl_k})] = 0$. $\qquad\square$

Let $\bar{L}$ be the Lipschitz constant of f. We are now ready to prove the convergence theorem of LPBNC under Assumption 1.

Theorem 4. *Let Assumption 1 hold and $\epsilon_{tol} = 0$. Suppose Algorithm 3 generates an infinite number of minor iterations after the $\bar{k}$-th major iteration. For every infinite subsequence $\mathcal{K} \subseteq \mathbb{N}$, if $B(\mathcal{K}) := \{l \in \mathcal{K} \mid a_l^{\bar{k}} < a^{th}(x^{\bar{k}}, I(\bar{k}, l))\}$ is a finite set, then*
(i) there exists $\{p_l\}_{l \in \mathcal{K}}$ such that $p_l \in P_{a_l^{\bar{k}}}(x^{\bar{k}})$ and $\|p_l - x^{\bar{k}}\|_{\infty} \xrightarrow{\mathcal{K}} 0$;
(ii) $0 \in \partial f(x^{\bar{k}})$.

Proof. (i) Let $\mathcal{K} \subseteq \mathbb{N}$ be an infinite subsequence and $B(\mathcal{K})$ be a finite set, then there exists $N_1 \in \mathcal{K}$ such that $a_l^{\bar{k}} \ge a^{th}(x^{\bar{k}}, I(\bar{k}, l))$ for all $l \ge N_1$ and $l \in \mathcal{K}$. By Lemma 4, $a_l^{\bar{k}}$ satisfies the required conditions in Lemma 3 for all $l \ge N_1$ and $l \in \mathcal{K}$. Suppose for contradiction that for all sequences $\{p_l\}_{l \in \mathcal{K}}$ such that $p_l \in P_{a_l^{\bar{k}}}(x^{\bar{k}})$, there exist $\epsilon > 0$ and $N_2 \in \mathcal{K}$ such that $\|p_l - x^{\bar{k}}\|_{\infty} \ge \epsilon$ for all $l \ge N_2$ and $l \in \mathcal{K}$. Then conclusion (43) can be applied with $(\bar{x}, x^*, a, \Delta, p, I)$ replaced by $(x^{\bar{k}}, x^{\bar{k},l}, a_l^{\bar{k}}, \Delta_l^{\bar{k}}, p_l, I(\bar{k}, l))$ for all $l \ge N_3 := \max\{N_1, N_2\}$ and $l \in \mathcal{K}$. For simplicity of notation we drop the superscript $\bar{k}$ and set $(x^l, a_l, \Delta_l, \rho_l) := (x^{\bar{k},l}, a_l^{\bar{k}}, \Delta_l^{\bar{k}}, \rho_l^{\bar{k}})$. We have

$$m(x^{\bar{k}}) - m(x^l) \ge [f(x^{\bar{k}}) - e_{a_l}(x^{\bar{k}})] \min\{\frac{\Delta_l}{\|x^{\bar{k}} - p_l\|_{\infty}}, 1\} \ge 0, \ \forall \ l \in \mathcal{K}_{\ge N_3}. \tag{58}$$

From Lemma 9(i) we have $m(x^{\bar{k}}) - m(x^l) \to 0$ as $l \to \infty$. Consequently,

$$[f(x^{\bar{k}}) - e_{a_l}(x^{\bar{k}})] \min\{\frac{\Delta_l}{\|x^{\bar{k}} - p_l\|_{\infty}}, 1\} \xrightarrow{\mathcal{K}} 0. \tag{59}$$

16

We first show that

$$[f(x^{\bar{k}}) - e_{a_l}(x^{\bar{k}})] \overset{\mathcal{K}}{\not\to} 0. \tag{60}$$

Suppose for contradiction that $[f(x^{\bar{k}}) - e_{a_l}(x^{\bar{k}})] \overset{\mathcal{K}}{\to} 0$. By Lemma 8, $\{a_l\}_{l \in \mathcal{K}}$ is bounded, and hence there exist a^* and $\bar{\mathcal{K}} \subseteq \mathcal{K}$ such that $a_l \overset{\bar{\mathcal{K}}}{\to} a^*$. As the mapping $e_{(\cdot)}(x^{\bar{k}})$ is continuous, we have $f(x^{\bar{k}}) = e_{a^*}(x^{\bar{k}})$ and equivalently $x^{\bar{k}} \in P_{a^*}(x^{\bar{k}})$. From the outer semicontinuity of the mapping $P_{(\cdot)}(x^{\bar{k}})$, there exist $\epsilon' < \epsilon$ and $N_4 \in \bar{\mathcal{K}}$ such that $\min_{p \in P_{a_l}(x^{\bar{k}})} \|x^{\bar{k}} - p\|_\infty \leq \epsilon'$, for all $l \in \bar{\mathcal{K}}_{\geq N_4}$. However, this cannot be true because $\epsilon' < \epsilon$, $\bar{\mathcal{K}} \subseteq \mathcal{K}$ and we have supposed that $\|p_l - x^{\bar{k}}\|_\infty \geq \epsilon$ for all $N_2 \leq l \in \mathcal{K}$, where p_l can be an arbitrary element of $P_{a_l}(x^{\bar{k}})$. We have finished showing (60) which together with (59) yields

$$\frac{\Delta_l}{\|x^{\bar{k}} - p_l\|_\infty} \overset{\mathcal{K}}{\to} 0. \tag{61}$$

Next we show

$$\{\|x^{\bar{k}} - p_l\|_\infty\}_{l \in \mathcal{K}} \text{ is bounded.} \tag{62}$$

By Definition 2.4, $f(p_l) \leq e_{a_l}(x^{\bar{k}}) \leq f(x^{\bar{k}})$, and therefore $p_l \in \text{lev}_{x^{\bar{k}}} f \subseteq \text{lev}_{x^0} f$. We also have $x^{\bar{k}} \in \text{lev}_{x^0} f$ which is bounded. Hence $\{\|x^{\bar{k}} - p_l\|_\infty\}_{l \in \mathcal{K}}$ is bounded above. We have supposed that $\|p_l - x^{\bar{k}}\|_\infty \geq \epsilon$ for all $N_2 \leq l \in \mathcal{K}$, hence $\{\|x^{\bar{k}} - p_l\|_\infty\}_{l \in \mathcal{K}}$ is bounded below and (62) is true. From (61) and (62) we have

$$\Delta_l \overset{\mathcal{K}}{\to} 0. \tag{63}$$

Then line 16 of Algorithm 3 must be executed infinite times, which implies that there exists an infinite subsequence $\mathcal{K}^* \subseteq \mathbb{N}$ such that $\rho_l < -\frac{1}{\min\{\Delta_l, 1\}}$ for all $l \in \mathcal{K}^*$ and $\mathcal{K}' := \mathcal{K} \cap \mathcal{K}^*$ is infinite. By the definition of ρ_l, the Lipschitz continuity of f, and the feasibility of x^l to problem (26) we have

$$m(x^{\bar{k}}) - m(x^l) < [f(x^{\bar{k}}) - f(x^l)] \min\{\Delta_l, 1\} \leq \bar{L}\|x^{\bar{k}} - x^l\|_\infty \min\{\Delta_l, 1\} \leq \bar{L}\Delta_l^2, \ \forall \, l \in \mathcal{K}'. \tag{64}$$

By (63) there exists $N_5 \in \mathcal{K}_{\geq N_2}$ such that $\Delta_l < \epsilon \leq \|p_l - x^{\bar{k}}\|_\infty$ for all $l \in \mathcal{K}_{\geq N_5}$. This implies $\min\{\frac{\Delta_l}{\|x^{\bar{k}} - p_l\|_\infty}, 1\} = \frac{\Delta_l}{\|x^{\bar{k}} - p_l\|_\infty}$ and from (58) we have

$$f(x^{\bar{k}}) - e_{a_l}(x^{\bar{k}}) \leq \frac{[m(x^{\bar{k}}) - m(x^l)]\|x^{\bar{k}} - p_l\|_\infty}{\Delta_l} \leq \bar{L}\|x^{\bar{k}} - p_l\|_\infty \Delta_l, \ \forall \, l \in (\mathcal{K}_{\geq N_6} \cap \mathcal{K}'), \tag{65}$$

where the last inequality follows from (64) and $N_6 := \max\{N_3, N_5\}$. From (62), (63), (65) and the fact that $\mathcal{K}' \subseteq \mathcal{K}$ we have

$$[f(x^{\bar{k}}) - e_{a_l}(x^{\bar{k}})] \overset{\mathcal{K}'}{\to} 0. \tag{66}$$

In (60) and its poof the $\mathcal{K}$ can be replaced by any infinite subsequence $\tilde{\mathcal{K}}$ such that $\|p_l - x^{\bar{k}}\|_\infty \geq \epsilon$ for all $l \in \tilde{\mathcal{K}}_{\geq N_2}$ with $p_l \in P_{a_l}(x^{\bar{k}})$. Consequently, (66) cannot be true and we have found a contradiction.

(ii) Finally, to see $0 \in \partial f(x^{\bar{k}})$, note $a_l(x^{\bar{k}} - p_l) \in \partial f(p_l)$ for all $p_l \in P_{a_l}(x^{\bar{k}})$ and $l \in \mathcal{K}$. By the outer semicontinuity of the proximal mapping and subdifferential, when $\|p_l - x^{\bar{k}}\|_\infty \overset{\mathcal{K}}{\to} 0$, as long as $\{a_l\}_{l \in \mathcal{K}}$ is bounded which is true from Lemma 8, we have $0 \in \partial f(x^{\bar{k}})$. $\qquad\square$

We denote the minor iterations between the k-th major iteration and the $(k+1)$-th major iteration $M(k) := \{0, 1, \cdots, l_k\}$ with $l_k = 0$ if there is no minor iteration in between. We will need the following assumption.

Assumption 2. *If there exists a sequence of indices $\{j_k\}_{k \in \mathcal{K}}$ with $j_k \in M(k)$ such that $G := \{k \in \mathbb{N} | \rho_{j_k}^k < -\frac{1}{\min\{\Delta_{j_k}^k, 1\}}\}$ is an infinite set then $\{k \in G | a_{j_k}^k < a^{th}(x^k, I(k, j_k))\}$ is a finite set.*

Theorem 5. *Let Assumption 1 and Assumption 2 hold and $\epsilon_{tol} = 0$. Suppose Algorithm 3 generates an infinite number of major iterations. For every subsequence $\mathcal{K} \subseteq \mathbb{N}$ and $\hat{x}$ such that $x^k \overset{\mathcal{K}}{\to} \hat{x}$, there exists an associated sequence of indices $\{i_k\}_{k \in \mathcal{K}}$ with $i_k \in M(k)$, such that if $E(\mathcal{K}) := \{k \in \mathcal{K} | a_{i_k}^k < a^{th}(x^k, I(k, i_k))\}$ is a finite set, then*
(i) there exists $\{p^k\}_{k \in \mathcal{K}}$ with $p^k \in P_{a_{i_k}^k}(x^k)$ such that $\|x^k - p^k\|_\infty \overset{\mathcal{K}}{\to} 0$;
(ii) $0 \in \partial f(\hat{x})$.

Proof. (i) Let $\mathcal{K} \subseteq \mathbb{N}$ be an infinite subsequence such that $x^k \xrightarrow{\mathcal{K}} \hat{x}$. Suppose for contradiction that for all sequences $\{p^k\}_{k \in \mathcal{K}}$ and $\{i_k\}_{k \in \mathcal{K}}$ with $i_k \in M(k)$, $p^k \in P_{a_{i_k}^k}(x^k)$ and $E(\mathcal{K})$ finite, there exist $\epsilon > 0$ and $M_1 \in \mathcal{K}$ such that $\|x^k - p^k\|_\infty \geq \epsilon$ for all $k \in \mathcal{K}_{\geq M_1}$. We first show that

$$[f(x^{\bar{k}}) - e_{a_{i_k}^k}(x^{\bar{k}})] \xrightarrow{\mathcal{K}} 0. \tag{67}$$

Suppose for contradiction that $[f(x^{\bar{k}}) - e_{a_{i_k}^k}(x^{\bar{k}})] \xrightarrow{\mathcal{K}} 0$. By Lemma 8, $\{a_{i_k}^k\}_{k \in \mathcal{K}}$ is bounded, and hence there exist a^* and $\bar{\mathcal{K}} \subseteq \mathcal{K}$ such that $a_{i_k}^k \xrightarrow{\bar{\mathcal{K}}} a^*$. As both $f(\cdot)$ and $e_{(\cdot)}(\cdot)$ are continuous, we have $f(\hat{x}) = e_{a^*}(\hat{x})$ and equivalently $\hat{x} \in P_{a^*}(\hat{x})$. From the outer semicontinuity of the mapping $P_{(\cdot)}(\cdot)$, there exist $\epsilon' < \epsilon$ and $M_2 \in \bar{\mathcal{K}}$ such that $\min_{p \in P_{a_{i_k}^k}(x^k)} \|x^k - p\|_\infty \leq \epsilon'$, for all $k \in \bar{\mathcal{K}}_{\geq M_2}$. However, this cannot be true because $\epsilon' < \epsilon$, $\bar{\mathcal{K}} \subseteq \mathcal{K}$ and we have supposed that $\|x^k - p^k\|_\infty \geq \epsilon$ for all $k \in \mathcal{K}_{\geq M_1}$, where p^k can be an arbitrary element of $P_{a_{i_k}^k}(x^k)$. We have finished showing (67).

Take $i_k \equiv l_k$ for all $k \in \mathcal{K}$. Since $E(\mathcal{K})$ is a finite set, there exists $M_3 \in \mathcal{K}$ such that $a_{l_k}^k \geq a^{th}(x^k, I(k, l_k))$ for all $k \in \mathcal{K}_{\geq M_3}$. By Lemma 4, $a_{l_k}^k$ satisfies the required conditions in Lemma 3 for all $k \in \mathcal{K}_{\geq M_3}$. Then conclusion (43) can be applied with $(\bar{x}, x^*, a, \Delta, p, I)$ replaced by $(x^k, x^{k,l_k}, a_{l_k}^k, \Delta_{l_k}^k, p^k, I(k, l_k))$ for all $k \in \mathcal{K}_{\geq M_4}$ where $M_4 := \max\{M_1, M_3\}$, i.e.

$$m(x^k) - m(x_{l_k}^k) \geq [f(x^k) - e_{a_{l_k}^k}(x^k)] \min\{\frac{\Delta_{l_k}^k}{\|x^k - p^k\|_\infty}, 1\} \geq 0, \ \forall \ k \in \mathcal{K}_{\geq M_4}.$$

From Lemma 9(ii) we have $m(x^k) - m(x^{k+1}) \to 0$ as $k \to \infty$. Consequently,

$$[f(x^k) - e_{a_{l_k}^k}(x^k)] \min\{\frac{\Delta_{l_k}^k}{\|x^k - p^k\|_\infty}, 1\} \xrightarrow{\mathcal{K}} 0. \tag{68}$$

From (67) and (68) we have

$$\frac{\Delta_{l_k}^k}{\|x^k - p^k\|_\infty} \xrightarrow{\mathcal{K}} 0. \tag{69}$$

Next we show

$$\{\|x^k - p^k\|_\infty\}_{k \in \mathcal{K}} \text{ is bounded.} \tag{70}$$

By Definition 2.4, $f(p^k) \leq e_{a_{l_k}^k}(x^k) \leq f(x^k)$, and therefore $p^k \in \text{lev}_{x^k} f \subseteq \text{lev}_{x^0} f$. We also have $x^k \in \text{lev}_{x^0} f$ which is bounded. Hence $\{\|x^k - p^k\|_\infty\}_{k \in \mathcal{K}}$ is bounded above. We have supposed that $\|x^k - p^k\|_\infty \geq \epsilon$ for all $k \in \mathcal{K}_{\geq M_1}$, hence $\{\|x^k - p^k\|_\infty\}_{k \in \mathcal{K}}$ is bounded below and (70) is true. From (69) and (70) we have

$$\Delta_{l_k}^k \xrightarrow{\mathcal{K}} 0. \tag{71}$$

Then line 16 of Algorithm 3 must be executed infinite times, which implies that there exists a subsequence $\mathcal{K}^* \subseteq \mathbb{N}$ such that $\mathcal{K}' := \mathcal{K} \cap \mathcal{K}^*$ is infinite and $\rho_{j_k}^k < -\frac{1}{\min\{\Delta_{j_k}^k, 1\}}$ for all $k \in \mathcal{K}^*$, with $j_k \in M(k)$. By the definition of $\rho_{j_k}^k$, the Lipschitz continuity of f, and the feasibility of x^{k,j_k} to problem (26) we have

$$\begin{aligned} m_{j_k}^k(x^k) - m_{j_k}^k(x^{k,j_k}) &< [f(x^k) - f(x^{k,j_k})] \min\{\Delta_{j_k}^k, 1\} \\ &\leq \bar{L} \|x^k - x^{k,j_k}\|_\infty \min\{\Delta_{j_k}^k, 1\} \\ &\leq \bar{L} \Delta_{j_k}^{k\ 2}, \ \forall \ k \in \mathcal{K}', \end{aligned} \tag{72}$$

where $m_{j_k}^k(\cdot) := m(\cdot; x^k, a_{j_k}^k, I(k, j_k))$.

We now show $\Delta_{j_k}^k \xrightarrow{\mathcal{K}'} 0$. Suppose for contradiction that $\{\Delta_{j_k}^k\}_{k \in \mathcal{K}'}$ is bounded away from 0. From (71) and the fact that $\mathcal{K}' \subseteq \mathcal{K}$, there exists $M_5 \in \mathcal{K}'$ such that

$$\Delta_{l_k}^k < \Delta_{j_{k+1}}^{k+1} \text{ for all } k \in \mathcal{K}'_{\geq M_5}. \tag{73}$$

18

From the update of trust region in Algorithm 3 we see that in minor iterations trust region radius is not increased and in major iterations trust region radius is increased under some conditions. Thus (73) implies that

$$\Delta_{j_{k+1}}^{k+1} \leq \Delta_0^{k+1} = \min\{\alpha_2 \Delta_{l_k}^k, \Delta_{\max}\} \leq \alpha_2 \Delta_{l_k}^k \text{ for all } k \in \mathcal{K}'_{\geq M_5}. \tag{74}$$

From (71), $\mathcal{K}' \subseteq \mathcal{K}$, (73) and (74), we have

$$\Delta_{j_k}^k \xrightarrow{\mathcal{K}'} 0. \tag{75}$$

We have finished showing $\Delta_{j_k}^k \xrightarrow{\mathcal{K}'} 0$ by reductio ad absurdum. Now for each j_k there exists $p_{j_k}^k \in P_{a_{j_k}^k}(x^k)$. If $\|p_{j_k}^k - x^k\|_\infty \xrightarrow{\mathcal{K}'} 0$ then we have found a sequence satisfying conclusion (i). Thus we suppose for contradiction that $\{\|p_{j_k}^k - x^k\|_\infty\}_{k \in \mathcal{K}'}$ is bounded away from 0. Consequently, there exists $M_6 \in \mathcal{K}'$ such that $\Delta_{j_k}^k < \epsilon \leq \|p_{j_k}^k - x^k\|_\infty$ for all $k \in \mathcal{K}'_{\geq M_6}$. This implies $\min\{\frac{\Delta_{j_k}^k}{\|x^k - p_{j_k}^k\|_\infty}, 1\} = \frac{\Delta_{j_k}^k}{\|x^k - p_{j_k}^k\|_\infty}$ for all $k \in \mathcal{K}'_{\geq M_6}$. Under Assumption 2, $\{k \in G | a_{j_k}^k < a^{th}(x^k, I(k, j_k))\}$ is a finite set, and therefore there exists $M_7 \in \mathcal{K}'$ such that Lemma 3 can be applied with $(\bar{x}, x^*, a, \Delta, p, I)$ replaced by $(x^k, x^{k,j_k}, a_{j_k}^k, \Delta_{j_k}^k, p_{j_k}^k, I(k, j_k))$ for all $k \in \mathcal{K}'_{\geq M_7}$, i.e.

$$f(x^k) - e_{a_{j_k}^k}(x^k) \leq \frac{[m_{j_k}(x^k) - m_{j_k}(x^{k,j_k})]\|x^k - p_{j_k}^k\|_\infty}{\Delta_{j_k}^k}, \ \forall \ k \in \mathcal{K}'_{\geq M_8}, \tag{76}$$

where $M_8 := \max\{M_6, M_7\}$. From (72) and (76) we have

$$f(x^k) - e_{a_{j_k}^k}(x^k) \leq \bar{L}\|x^k - p_{j_k}^k\|_\infty \Delta_{j_k}^k, \ \forall \ k \in \mathcal{K}'_{\geq M_8}. \tag{77}$$

The p^k in (70) and its proof can be replaced by $p_{j_k}^k$ and thus $\{\|x^k - p_{j_k}^k\|_\infty\}_{k \in \mathcal{K}'}$ is bounded. This together with (75) and (77) yields $f(x^k) - e_{a_{j_k}^k}(x^k) \xrightarrow{\mathcal{K}'} 0$. We can easily check that the $\mathcal{K}$ in (67) and its proof can be replaced by $\mathcal{K}'$ with i_k replaced by j_k and p^k replaced by $p_{j_k}^k$. Hence we have $f(x^k) - e_{a_{j_k}^k}(x^k) \xrightarrow{\mathcal{K}'} 0$. We have found a contradiction and conclusion (i) holds true.

(ii) To see $0 \in \partial f(\hat{x})$, note $a_{i_k}^k(x^k - p^k) \in \partial f(p^k)$ for all $p^k \in P_{a_{i_k}^k}(x^k)$ and $k \in \mathcal{K}$. By the outer semicontinuity of the proximal mapping and subdifferential, when $x^k \xrightarrow{\mathcal{K}} \hat{x}$ and $\|p^k - x^k\|_\infty \xrightarrow{\mathcal{K}} 0$, as long as $\{a_{i_k}^k\}_{k \in \mathcal{K}}$ is bounded which is true from Lemma 8, we have $0 \in \partial f(\hat{x})$. $\qquad \square$

6 Numerical experiments

In this section we report some preliminary numerical results on implementations of LPBC and LPBNC. Here our goal is to provide a proof of principle only. For nonconvex examples we demonstrate that the conditions stated in the convergence theorems can be satisfied. The two algorithms were programmed in MATLAB R2012b in a computer with 3.40 GHz CPU and 8 GB RAM. We used the CPLEX connector (V12.5.1) for MATLAB to solve the linear subproblems. Specifically, problem (9) was programmed and solved by the CPLEX Class API and problem (26) was solved by the toolbox function cplexlp. CPLEX automatically chooses from primal simplex, dual simplex and barrier optimizers to solve a given linear programming problem. In our implementations we found that no instances of the linear subproblems were solved by barrier optimizer. The implementation of our algorithms requires high accuracy of the solution of linear subproblems. Hence the CPLEX tolerances of optimality and feasibility are crucial to the performance of LPBC and LPBNC. As we see from the algorithms, both the stopping criterion and definition of ρ are dependent on the model reduction, $f(\bar{x}) - z^*$; if z^* provided by CPLEX solver is slightly bigger than the actual optimal value of (26), then the model reduction can be significantly inaccurate. In fact, we observed instances of negative model reduction when we use the default tolerances of optimality and feasibility in CPLEX. To prevent the occurrence of such cases, we set both the optimality and feasibility tolerances to 10^{-9}, the least value in CPLEX.

In our experiments, the choice of the initial trust region radius Δ_0^0 can significantly change the performance of our algorithm on some problems. In most trust region methods, choosing the initial trust region radius is an import issue, as stated in the monograph [3, page 784] "one very often has to resort to some heuristic to choose Δ_0 on the basis of other initial information." Nonetheless, [3] suggested several strategies of initializing trust region and we adopted the following two choices, $\Delta_0^0 = 1$ and $\Delta_0^0 = \frac{1}{10}\|s^0\|$, where s^0 is an element of $\partial f(x^0)$. Settings for other trust region related parameters in both LPBC and LPBNC are $\eta_1 = 10^{-4}$, $\eta_3 = 0.4$, $\alpha_1 = 0.25$, $\alpha_2 = 2$, and $\Delta_{\max} = 1000$.

6.1 Convex Examples

Table 2 presents the results for the tested convex problems listed in Table 1, where problems 1 to 15 are taken from section 3 of [18] and problems 16 to 20 are the problems 2.1 to 2.5 of [11]. For all convex and nonconvex problems in our tests we use the initial points provided in the associated references. The following abbreviations

Table 1: Tested convex problems

No.	Problem	Dimension	Optimal Value
1	CB2	2	1.9522245
2	CB3	2	2
3	DEM	2	-3
4	QL	2	7.2
5	LQ	2	-1.4142136
6	Mifflin1	2	-1
7	Wolfe	2	-8
8	Rosen	4	-44
9	Shor	5	22.600162
10	Maxquad	10	-0.8414083
11	Maxq	20	0
12	Maxl	20	0
13	Goffin	50	0
14	MXHILB	50	0
15	L1HILB	50	0
16	Generalization of MAXQ	100	0
17	Generalization of MXHILB	100	0
18	Chained LQ	100	$-99\sqrt{2}$
19	Chained CB3 I	100	198
20	Chained CB3 II	100	198

are used in Table 2.

f_{val}	minimal function value returned by the algorithm,
nf	number of function evaluations used by the algorithm,
k	number of major iterations,
L	number of minor iterations,
time	elapsed CPU time,
t-CPX	sum of CPU time by CPLEX solver,
Δ	final value of trust region radius,
sh	number of times that trust region radius is decreased,
pr	number of times that primal simplex method is chosen by CPLEX,
dual	number of times that dual simplex method is chosen by CPLEX.

In LPBC, we set $\epsilon_{tol} = 10^{-6}$ and $T = 30$. For problems 1 - 14 we initialize trust region radius by 1; for problems 15 - 20 we initialize trust region radius by $\frac{1}{10}\|s^0\|$. From Table 2 we see that LPBC returned optimal values to accuracy 10^{-6} for 14 problems, 10^{-4} for 4 problems and 2×10^{-4} for 2 problems. For all problems except the last three, CPLEX consumed negligible time to solve all the linear subproblems. The possibility that trust region is shrunk (the value sh/L) is very small for the majority of the tested problems. We also see that most of the linear subproblems were solved by dual simplex method.

6.2 Nonconvex Examples

The tested nonconvex problems are listed in Table 3 where problems 1 to 7 are taken from section 3 of [18] and problems 8 to 12 are the problems 2.6 to 2.10 in [11].

In LPBNC, we set $\gamma = 2$ and $\epsilon_{tol} = 10^{-5}$. Apart from the initial trust region radius Δ_0^0, another parameter that can cause dramatic changes in the performance of LPBNC is the backtrack parameter β. Results with two settings of these parameters are listed in Table 4 and 5 below. We use the same abbreviations as in Table 2. Additionally, we list the difference of f_{val} and optimal value (error), the number of function evaluations

Table 2: Results for convex problems

No.	f_{val}	nf	k	L	time	t-CPX	Δ	sh	pr	dual
1	1.952225451	16	10	5	0.1428	0	0.341842126	0	0	16
2	2	3	1	1	0.0977	0	0.5	1	0	3
3	-3	8	6	1	0.1097	0	2.039607805	0	2	6
4	7.20000069	16	10	5	0.1322	0	0.141421356	2	0	16
5	-1.41421274	18	14	3	0.1305	0	1.13137085	0	3	15
6	-0.999999683	28	12	15	0.1562	0	0.03125	6	5	23
7	-8	5	3	1	0.1021	0	2	1	0	5
8	-43.99998585	54	24	29	0.2361	0	0.25	4	1	53
9	22.60018019	55	22	32	0.2356	0	0.25	1	0	55
10	-0.841407474	220	34	185	0.7292	0	0.25	1	0	220
11	4.06847E-07	249	116	132	0.8042	0	2	3	0	249
12	0	36	16	19	0.1735	0	3.2	2	7	29
13	0	51	48	2	0.2316	0.016	4	0	1	50
14	2.35525E-07	15	10	4	0.1290	0	8.158764414	2	3	12
15	2.08721E-06	27	15	11	0.2028	0	0.13964447	6	1	26
16	4.21692E-07	1361	583	777	4.5585	0.079	40	2	1	1360
17	9.97664E-07	25	15	9	0.1571	0	2.045863824	5	3	22
18	-140.0070287	1185	91	1093	8.4852	3.661	0.124058958	3	0	1185
19	198.000171	1437	132	1304	10.4948	5.109	0.13978045	5	2	1435
20	198.0000905	35612	199	35412	285.4227	150.166	0.13978045	5	2	35610

Table 3: Tested nonconvex problems

No.	Problem	Dimension	Optimal Value
1	Crescent	2	0
2	Mifflin2	2	-1
3	Colville 1	5	-32.348679
4	HS78	5	-2.9197004
5	El-Attar	6	0.5598131
6	Gill	10	9.7857721
7	Steiner 2	12	16.703838
8	Active Faces	50	0
9	Brown 2	50	0
10	Chained Mifflin2	50	-34.795
11	Chained Crescent I	50	0
12	Chained Crescent II	50	0

in the backtrack process divided by the total number of function evaluations (pb), the number of subgradient evaluations (se), the final value for a, the final value for $a^{\min}$ and the number of times that a is updated (au). Subgradient evaluations do not happen in the backtrack process and hence the total number of subgradient evaluations is not equal to that of function evaluations. We found that all instances of LP subproblems were solved by dual simplex method in our tests for nonconvex problems. We see in Table 4 problem 8 and 9 have

Table 4: Results for nonconvex problems with $\Delta_0^0 = 1,\ \beta = 0.7$

No.	error	nf	pb	se	k	L	time	t-CPX	Δ	sh	a	$a^{\min}$	au
1	0.000257	46	30.43	32	19	12	1.186	0	0.0625	3	0.90002	0.89339	3
2	0	20	10.00	18	4	13	0.260	0	0.03125	4	0	0	0
3	6.28E-07	46	19.57	37	14	22	0.561	0	0.125	3	0	0	0
4	6.5E-05	81	2.47	79	30	48	0.940	0	0.125	4	50.37528	48.16938	5
5	9.25E-07	99	9.09	90	35	54	1.214	0.031	0.125	8	0.40795	0.39540	9
6	0.0002196	205	1.95	201	30	170	4.602	0.076	0.015625	4	6.769637	6.769637	5
7	0.0001409	123	16.26	103	36	66	1.116	0.03	0.03125	3	0	0	0
8	0	2	0.00	2	1	0	0.278	0	2	0	0.118057	0.118057	1
9	0	2	0.00	2	1	0	0.121	0	2	0	0	0	0
10	4.05E-05	11549	69.81	3487	85	3401	345.646	16.931	0.003906	5	0	0	0
11	7.29E-06	5743	38.22	3548	189	3358	126.861	9.896	0.003906	5	0	0	0
12	1.01E-05	1637	54.31	748	72	675	15.801	1.199	0.000977	9	0	0	0

error 0 with two function evaluations. These two instances are accidental as we found that the optimal solution of the first linear subproblem (26) is already the minimizer of the objective function for these two problems given that we set $\Delta_0^0 = 1$ and we use the default initial points. We note that a huge negative error occurred

Table 5: Results for nonconvex problems with $\Delta_0^0 = \frac{1}{10}\|s^0\|,\ \beta = 0.8$

No.	error	nf	pb	se	k	L	time	t-CPX	Δ	sh	a	$a^{\min}$	au
1	0.0002435	40	37.50	25	19	5	0.311	0	0.10607	3	0.375	0.363	3
2	1.095E-05	20	5.00	19	6	12	0.247	0.016	0.03542	4	0	0	0
3	0.0001802	49	28.57	35	15	19	0.390	0	0.02140	6	0	0	0
4	-6.636E+19	15	0.00	15	14	0	0.215	0	1000	0	3.78E+12	2.81E+12	13
5	1.587E-05	437	48.97	223	105	117	2.738	0.015	0.13473	28	0.196	0.163	49
6	0.0002527	458	26.64	336	75	260	4.295	0.11	0.04621	6	7.995	0	15
7	0.0001122	118	14.41	101	33	67	1.071	0.03	0.06833	3	0	0	0
8	0.0003365	1028	24.42	777	73	703	18.732	1.076	0.00693	16	0.027472	0.027465	46
9	3.334E-06	657	34.70	429	36	392	9.414	0.712	0.00136	12	0.005	0.003	12
10	2.918E-05	22741	81.43	4222	88	4133	369.968	18.958	0.00544	7	0	0	0
11	6.576E-06	9530	51.29	4642	180	4461	201.156	14.651	0.00238	7	2.485	0	4
12	8.237E-06	1801	64.02	648	73	574	21.416	1.587	0.00059	9	0	0	0

in problem 4 in Table 5 because the problem HS78 is actually unbounded below and its optimal value in Table 3 is a local minimum. From Table 4 and 5 we see that the proportion of function evaluations in backtrack process can be very high. Problem 10, which has the biggest number of function evaluations also yields the highest possibility to backtrack. For all the problems, solving LP subproblems took a small amount of time. objective function. This special case appeared in all the tested Fuse errier Polynomials and Active Faces. We ran LPBNC on these problems with all choices of dimension $n \in \{2, 10, 10^2, 10^3, 10^4, 10^5, 10^6\}$. The starting point is $\vec{x}^0 = [1, 1, \ldots, 1]$, initial trust region radius $\Delta_0^0 = 1$. All theses instances terminated with stopping criterion reached, $f_{val} = 0$, and $nf = 3$.

7 Concluding remarks

We present a version of bundle method with the unusual feature of using only an LP solver as its algorithmic engine. We study the properties of the linear model and expressed its model reduction. The optimal solution of our linear subproblem is not unique in contrast to the case of quadratic subproblem. However, no significant information is lost in order to ensure convergence of the algorithm. We use a local convexificaton with the deletion of some cutting planes at the end of a major iteration. Preliminary numerical experiments show that the algorithm is reliable and efficient for solving convex problems. For functions that are locally Lipschitz continuous and prox-regular we show that upon successful convexification the algorithm can converge to a fixed point of the proximal mapping. Numerical results of nonconvex problems suggest that with insignificant time spent on solving LP subproblems, a big portion of function evaluations can be consumed in the backtrack process. Improvements such as incorporation of a line search can be further studied in the future in order to increase efficiency to enable the solution of large-scale problems.

References

[1] Bagirov, A.M., Jin, L., Karmitsa, N., Al Nuaimat, A., Sultanova, N.: Subgradient method for nonconvex nonsmooth optimization. Journal of Optimization Theory and Applications pp. 1–20 (2013). URL http://link.springer.com/article/10.1007/s10957-012-0167-6

[2] Burke, J., Lewis, A., Overton, M.: A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. SIAM Journal on Optimization **15**(3), 751–779 (2005)

[3] Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust Region Methods. Society for Industrial and Applied Mathematics (2000). DOI 10.1137/1.9780898719857. URL http://epubs.siam.org/doi/abs/10.1137/1.9780898719857

[4] Dempe, S., Bard, J.F.: Bundle trust-region algorithm for bilinear bilevel programming. Journal of Optimization Theory and Applications **110**(2), 265–288 (2001)

[5] Eberhard, A.: Prox-regularity and subjets. In: A. Rubinov, B. Glover (eds.) Optimization and Related Topics, *Applied Optimization*, vol. 47, pp. 237–313. Springer US (2001). DOI 10.1007/978-1-4757-6099-6_14. URL http://dx.doi.org/10.1007/978-1-4757-6099-6_14

[6] Feltenmark, S., Kiwiel, K.C.: Dual applications of proximal bundle methods, including lagrangian relaxation of nonconvex problems. SIAM Journal on Optimization **10**(3), 697–721 (2000)

[7] Fuduli, A., Gaudioso, M., Nurminski, E.: A splitting bundle approach for non-smooth nonconvex minimization. Optimization pp. 1–21 (2013). DOI 10.1080/02331934.2013.840625. URL http://dx.doi.org/10.1080/02331934.2013.840625

[8] Hare, W., Sagastizábal, C.: A redistributed proximal bundle method for nonconvex optimization. SIAM Journal on Optimization **20**(5), 2442–2473 (2010). DOI 10.1137/090754595. URL http://dx.doi.org/10.1137/090754595

[9] Hiriart-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms II, *Grundlehren der mathematischen Wissenschaften*, vol. 306. Springer Berlin Heidelberg (1993). DOI 10.1007/978-3-662-06409-2

[10] Jeyakumar, V., Yang, X.: Convex composite minimization with $c_{1,1}$ functions. Journal of Optimization Theory and Applications **86**(3), 631–648 (1995)

[11] Karmitsa, N.: Test problems for large-scale nonsmooth minimization. Reports of the Department of Mathematical Information Technology, Series B, Scientific computing, No. B **4** (2007)

[12] Kiwiel, K.C.: An aggregate subgradient method for nonsmooth convex minimization. Mathematical Programming **27**(3), 320–341 (1983)

[13] Kiwiel, K.C.: Methods of descent for nondifferentiable optimization, vol. 1133. Springer-Verlag Berlin (1985)

[14] Kiwiel, K.C.: Proximity control in bundle methods for convex nondifferentiable minimization. Mathematical Programming **46**(1-3), 105–122 (1990)

[15] Lemarechal, C.: An extension of davidon methods to non differentiable problems. In: M. Balinski, P. Wolfe (eds.) Nondifferentiable Optimization, *Mathematical Programming Studies*, vol. 3, pp. 95–109. Springer Berlin Heidelberg (1975). DOI 10.1007/BFb0120700. URL http://dx.doi.org/10.1007/BFb0120700

[16] Lewis, A., Overton, M.: Nonsmooth optimization via quasi-newton methods. Mathematical Programming **141**(1-2), 135–163 (2013). DOI 10.1007/s10107-012-0514-2. URL http://dx.doi.org/10.1007/s10107-012-0514-2

[17] Linderoth, J., Wright, S.: Decomposition algorithms for stochastic programming on a computational grid. Computational Optimization and Applications **24**(2-3), 207–250 (2003). DOI 10.1023/A:1021858008222. URL http://dx.doi.org/10.1023/A%3A1021858008222

[18] Lukšan, L., Vlček, J.: Test problems for nonsmooth unconstrained and linearly constrained optimization. Tech. Rep. 798, INSTITUTE OF COMPUTER SCIENCE, ACADEMY OF and SCIENCES OF and THE CZECH and REPUBLIC (2000)

[19] Lukšan, L., Vlcek, J.: Variable metric method for minimization of partially separable nonsmooth functions. Pacific Journal on Optimization **2**, 59–70 (2006)

[20] Qi, L., Sun, J.: A nonsmooth version of newton's method. Mathematical Programming **58**(1-3), 353–367 (1993). DOI 10.1007/BF01581275. URL http://dx.doi.org/10.1007/BF01581275

[21] Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM Journal on Control and Optimization **14**(5), 877–898 (1976)

[22] Rockafellar, R.T., Wets, R.J.B.: Variational Analysis: Grundlehren Der Mathematischen Wissenschaften, vol. 317. Springer (1998)

[23] Schramm, H., Zowe, J.: A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. SIAM Journal on Optimization **2**(1), 121–152 (1992). DOI 10.1137/0802008. URL http://epubs.siam.org/doi/abs/10.1137/0802008

[24] Shor, N.: Application of the Gradient Method for the solution of Network Transportation Problems (1962)

[25] Shor, N.: Minimization Methods for Non-Differentiable Functions, *Springer Series in Computational Mathematics*, vol. 3. Springer Berlin Heidelberg (1985)

[26] Tang, C.m., Liu, S., Jian, J.b., Li, J.l.: A feasible sqp-gs algorithm for nonconvex, nonsmooth constrained optimization. Numerical Algorithms **65**(1), 1–22 (2014). DOI 10.1007/s11075-012-9692-5. URL http://dx.doi.org/10.1007/s11075-012-9692-5

[27] Vaz, A.I.F., Vicente, L.N.: A particle swarm pattern search method for bound constrained global optimization. Journal of Global Optimization **39**(2), 197–219 (2007)

[28] Wolfe, P.: A method of conjugate subgradients for minimizing nondifferentiable functions. In: M. Balinski, P. Wolfe (eds.) Nondifferentiable Optimization, *Mathematical Programming Studies*, vol. 3, pp. 145–173. Springer Berlin Heidelberg (1975). DOI 10.1007/BFb0120703. URL http://dx.doi.org/10.1007/BFb0120703

A   Appendix

A.1   Proof for Theorem 2

In order to prove Theorem 2 we need the following preliminaries. Let X denote a space with $\|\cdot\|$ its norm. For every $A \subseteq X$, $\alpha > 0$ and $x^* \in X^*$ we define

$$SL(x^*, A, \alpha) = \{x \in A \mid \langle x, x^* \rangle > S(A, x^*) - \alpha\},$$

where $S(A, x^*) := \sup \{\langle x^*, x \rangle \mid x \in A\}$. We will denote a linear function on X by x^* or $\langle x^*, \cdot \rangle$ and their action an element x by $x^*(x)$ or $\langle x^*, x \rangle$. Denote the set of all strongly exposed points of a set C by $\exp C$ and the set of all extremal points by $\text{ext}\, C$. We have for any closed, proper convex function h

$$\text{epi}\, h = \overline{\text{co}}\,(\text{ext}\,\text{epi}\, h) + \mathbf{R}_+ = \overline{\text{co}}\,(\exp \text{epi}\, h) + \mathbf{R}_+.$$

Suppose $(x, h(x)) \in \text{ext}\,\text{epi}\, h$ then $h(x) = g(x)$. In this section we use the following definition of proximal mapping $P_\lambda (f)(x) := \arg\min \{f(\cdot) + \frac{1}{2\lambda}\|x - \cdot\|^2\}$ and $\overline{r}(f)$ is the associated prox-threshold. It is known that if $f : \mathbf{R}^n \to \mathbf{R}_{+\infty}$ is quadratically minorized by $\alpha - r\|\cdot\|^2$ and $x \in \text{dom}\, f$ with $0 < \lambda < \overline{r}(f, x)$ then for each $\beta > f(x)$ we have $\exists \delta > 0$ such that

$$\|z - y\| \leq 2\lambda \frac{(\beta - \alpha + \frac{r}{2})(1 - 2\delta)}{(1 - 2(\delta + r\lambda))} := M_\lambda \,\forall\, y \in P_\lambda(f)(z) \,\forall\, z \in B_\delta(x). \tag{78}$$

Lemma 10. *Suppose $h : D \to \mathbf{R}$ is an lower bounded, real–valued lower semi–continuous convex function defined on a bounded closed convex set $D \subseteq \mathbf{R}^n$, with interior. Then there exists a $(y^*, -1)$ strongly exposing $\text{epi}\, h$ at $(\bar{y}, h(\bar{y}))$ if and only if $h - y^*$ achieves a strict minimum on D at $\bar{y}$. That is, $\arg\min [h - y^*] = \{y\}$ or $0 \in \partial [h - y^*](y)$ has a unique solution y.*

Proof. Indeed if $h - y^*$ achieves a strict minimum $\bar{\beta} := \min_{y \in D}\{h(y) - \langle y^*, y \rangle\}$ then we must have the lower level sets

$$L(\beta) := \{y \in D \mid \beta > h(y) - \langle y^*, y \rangle\}$$

satisfying $\cap_{\beta < \bar{\beta}} L(\beta) = \{\bar{y}\}$. Otherwise there would exists $y_m \in L(\beta_m)$ for $\beta_m \downarrow \bar{\beta}$ converging to $y \neq \bar{y}$ with $y \in \cap_{\beta < \bar{\beta}} L(\beta)$ and by the lower continuity of h we have

$$\liminf_m \beta_m = \bar{\beta} \geq \liminf_m (h(y_m) - \langle y^*, y_m \rangle) \geq h(y) - \langle y^*, y \rangle$$

implying $y \in \arg\min (h - y^*)$. This contradicts the uniqueness of the minimizer. Thus the slices

$$L(\bar{\beta} - \delta) = \{y \in D \mid h(\bar{y}) - \langle y^*, \bar{y} \rangle + \delta > h(y) - \langle y^*, y \rangle\} \to \{\bar{y}\}$$

as $\delta \downarrow 0$. Observe that $S(\text{epi}\, h, (-y^*, 1)) = \langle (\bar{y}, h(\bar{y})), (-y^*, 1) \rangle$ and so

$$L(\bar{\beta} + \delta) = \{y \in D \mid \langle (y, h(y)), (-y^*, 1) \rangle > S(\text{epi}\, h, (-y^*, 1)) - \delta\}.$$

Finally note that on $\text{epi}\, h$ we have

$$SL((-y^*, 1), \text{epi}\, h, \delta) = \{(y, \alpha) \in \text{epi}\, h \mid \langle (y, \alpha), (-y^*, 1) \rangle > S(\text{epi}\, h, (-y^*, 1)) - \delta\} \tag{79}$$

and $\text{diam}\, SL((-y^*, 1), \text{epi}\, h, \delta) \leq \text{diam} L(\bar{\beta} + \delta) \times \delta \to 0$ as $\delta \downarrow 0$.

If $(y^*, -1)$ strongly exposing $\text{epi}\, h$ at $(\bar{y}, h(\bar{y}))$ then (79) defines a slice whose diameter tends to zero. It then holds that the projection of this onto D also has a diameter tending to zero from which it follows that $\text{diam}\, L(\bar{\beta} + \delta) \to 0$ as $\delta \downarrow 0$. Clearly we have then $\cap_{\beta < \bar{\beta}} L(\beta) = \{\bar{y}\}$ and so $h - y^*$ achieves a strict mimimum $\bar{\beta}$ at $\bar{y}$. $\qquad\square$

Lemma 11. *Suppose $h : D \to \mathbf{R}$ is an lower bounded, real–valued lower semi–continuous, convex function defined on a bounded closed convex set $D \subseteq \mathbf{R}^n$, with interior. Suppose $(x, h(x))$ is an extremal point of $\text{epi}\, h$ and suppose $\{(x_m, \alpha_m)\}_{m=0}^\infty \subseteq \text{epi}\, h$ with $x_m = \sum_{i=0}^n \lambda_i^m x_i^m \to x$ and $\alpha_m = \sum_{i=0}^n \lambda_i^m h(x_i^m) \to h(x)$. Then either $\lambda_i^m \to 0$ or $x_i^m \to x$ and hence for some i we have $x_i^m \to x$.*

Proof. We now claim that either $\lambda_i^m \to 0$ or $x_i^m \to x$ for otherwise by a compactness argument (on $[0, 1] \times D$) we could extract a convergent pair of sub-sequences such that after renumbering we would have $\lambda_i^m \to \lambda \neq 0$ and $x_i^m \to x' \neq x$ (note that $\lambda \in [0, 1]$ and $x' \in D$). Then as $\lambda_i^m x_i^m \to \lambda x'$ converges as does $\sum_{i=0}^n \lambda_i^m x_i^m \to x$ we have convergence of $\sum_{j \neq i} \left(\frac{\lambda_j^m}{1 - \lambda_i^m}\right) x_j^m$ to $x'' \in D$ via

$$\lambda_i^m x_i^m + (1 - \lambda_i^m) \sum_{j \neq i} \left(\frac{\lambda_j^m}{1 - \lambda_i^m}\right) x_j^m \to \lambda x' + (1 - \lambda) x'' = x.$$

It is not possible for $\lambda = 1$ because this implies $x' = x$. Next note that

$$\lambda_i^m h(x_i^m) + (1 - \lambda_i^m) \sum_{j \neq i} \left(\frac{\lambda_j^m}{1 - \lambda_i^m} \right) h(x_j^m) \to h(x) \tag{80}$$

and by lower semi–continuity of h we have $\liminf_m h(x_i^m) \geq h(x')$. If we take a subsequence along which the limit infimum $\liminf_m h(x_i^m)$ is achieved then (80) implies $\left\{ \lim_{m_k} \sum_{j \neq i} \left(\frac{\lambda_j^{m_k}}{1 - \lambda_i^{m_k}} \right) h(x_j^{m_k}) \right\}$ converges along this subsequence as well. Let

$$\liminf_{m_k} \sum_{j \neq i} \left(\frac{\lambda_j^{m_k}}{1 - \lambda_i^{m_k}} \right) (x_j^{m_k}, h(x_j^{m_k})) = (x'', \alpha'') \in \text{epi } h,$$

with $h(x'') \leq \alpha''$ by the definition of epi h. Then from (80) and the lower semi–continuity of h again it follows that

$$\lambda h(x') + (1 - \lambda) h(x'') \leq h(x)$$

Thus there exists (x', α'), $(x'', \alpha'') \in \text{epi } h$ such that $\lambda(x', \alpha') + (1 - \lambda)(x'', \alpha'') = (x, h(x))$ with $0 < \lambda < 1$ contradicting the assumption that $(x, h(x))$ is an extremal point of epi h.

As either $\lambda_i^m \to 0$ or $x_i^m \to x$ it follows that for some i we have $x_i^m \to x$ since $\sum_{i=0}^n \lambda_i^m = 1$ precludes all λ_i^m from tending to zero. $\square$

Proposition 2. *Suppose $h : D \to \mathbf{R}$ is an lower bounded, real–valued lower semi–continuous, convex function defined on a bounded closed convex set $D \subseteq \mathbf{R}^n$, with interior. In addition suppose the strongly exposed points on h are <u>dense on</u> the boundary of epi $h|_C$ where $C \subseteq D$ and C is closed with $\text{int} C \neq \emptyset$. Define g via epi $g := \overline{\exp \text{ epi } h|_C}$ then*

$$h(x) = g(x) \quad \text{for } x \in C.$$

Proof. On C take a $(x, h(x)) \in \exp \text{ epi } h$ and any supporting hyperplane generated by $(-y^*, 1)$. By definition of exposedness we have

$$\text{epi } h \cap \{(y, \alpha) \mid \langle (-y^*, 1), (x, h(x)) \rangle \geq \langle (-y^*, 1), (y, \alpha) \rangle\} = \{(x, h(x))\}.$$

Now

$$
\begin{aligned}
\text{epi } h|_C \;&=\; \text{co } \exp \text{ epi } h|_C = \cap_{\substack{(x, h(x)) \in \exp \text{ epi } h \\ x \in C}} \{(y, \alpha) \mid \langle (-y^*, 1), (x, h(x)) \rangle \leq \langle (-y^*, 1), (y, \alpha) \rangle\} \cap [C \times \mathbf{R}] \\
&\subseteq\; \text{epi } h.
\end{aligned}
$$

Now suppose there exists $(y, \alpha) \in \text{co } \exp \text{epi} h|_C \cap (\text{epi } g)^c$. Using the density of $\exp \text{ epi } h|_C$ we have

$$g(y) = \liminf_{\substack{(x, h(x)) \in \exp \text{ epi } h \\ x \in C,\ x \to y}} h(x).$$

As $(x, h(x)) \in \exp \text{ epi } h \subseteq \text{epi} h$ with $x \in C$ and $(y, \alpha) \in \text{co } \exp \text{ epi } h|_C$ implies $\langle (-y^*, 1), (x, h(x)) \rangle \leq \langle (-y^*, 1), (y, \alpha) \rangle$ we have

$$\liminf_{\substack{(x, h(x)) \in \exp \text{ epi } h \\ x \in C,\ x \to y}} \langle (-y^*, 1), (x, h(x)) \rangle = \langle (-y^*, 1), (y, g(y)) \rangle \leq \langle (-y^*, 1), (y, \alpha) \rangle.$$

Thus $g(y) \leq \alpha$ or $(y, \alpha) \in \text{epi } g$, a contradiction. Hence co $\exp \text{epi } h|_C \cap [C \times \mathbf{R}] \subseteq \text{epi } g$ giving the equality epi $h|_C =$ co $\exp \text{ epi } h|_C \cap [C \times \mathbf{R}] = \text{epi } g$. $\square$

Corollary 1. Suppose $g : D \to \mathbf{R}$ is an lower bounded, real–valued continuous function defined on a bounded closed set $D \subseteq \mathbf{R}^n$, with interior. If the strongly exposed points of epi g are dense in the boundary of epi g then $g(x) = \text{co } g(x)$ for all $x \in D$ and hence g is the restriction to D of a convex function. In particular this is true when $0 \in \partial [g - y^*](y)$ has a unique solution for all $y^* \in B_1(0)$ such that $(-y^*, 1)$ supports epi g.

Proof. Let $h := \text{co epi } g = \text{epi co } g$ which has a convex domain co D with $\text{intco } D \supseteq \text{int}D \neq \emptyset$. Clearly a strongly exposed point of h must be a strongly exposed point of g. We claim the converse is true. Let $(x, g(x))$ be strongly exposed in epi g. Suppose $(x, h(x))$ is not strongly exposed in epi h then there exists $(-y^*, 1)$ supporting epi h at $(x, h(x))$ such that for some $\delta > 0$ and all $\varepsilon > 0$ there exists $(y^\varepsilon, h(y^\varepsilon)) \in \text{epi } h$ with $\|(y^\varepsilon, h(y^\varepsilon)) - (x, h(x))\| \geq \delta > 0$ and

$$\langle (-y^*, 1), (x, h(x)) \rangle + \varepsilon \geq \langle (-y^*, 1), (y, h(y^\varepsilon)) \rangle.$$

As $g(x) \geq h(x)$ and $h(y^\varepsilon) = \sum_{i=0}^n \lambda_i^\varepsilon g(x_i^\varepsilon)$ for some $\sum_{i=0}^n \lambda_i^\varepsilon = 1$, $\lambda_i^\varepsilon \geq 0$ and $y^\varepsilon = \sum_{i=0}^n \lambda_i^\varepsilon x_i^\varepsilon$ we have

$$\langle (-y^*, 1), (x, g(x)) \rangle + \varepsilon \geq \sum_{i=0}^n \lambda_i^\varepsilon \langle (-y^*, 1), (x_i^\varepsilon, g(x_i^\varepsilon)) \rangle$$

$$\text{or} \qquad \sum_{i=0}^n \lambda_i^\varepsilon \langle (-y^*, 1), (x, g(x)) - (x_i^\varepsilon, g(x_i^\varepsilon)) \rangle \quad \geq \quad -\varepsilon.$$

As $(-y^*, 1)$ supporting epi h we have $\langle (-y^*, 1), (x, g(x)) - (x_i^\varepsilon, g(x_i^\varepsilon)) \rangle \leq 0$ for all i. Thus for all i with $\lambda_i^\varepsilon > 0$ we have

$$\lambda_i^\varepsilon \langle (-y^*, 1), (x, g(x)) - (x_i^\varepsilon, g(x_i)) \rangle$$

$$\geq \quad \sum_{i=0}^n \lambda_i^\varepsilon \langle (-y^*, 1), (x, g(x)) - (x_i^\varepsilon, g(x_i)) \rangle \geq -\varepsilon.$$

We have $\langle (-y^*, 1), (x, g(x)) - (x_i^\varepsilon, g(x_i^\varepsilon)) \rangle \geq -\varepsilon/\lambda_i^\varepsilon$ for all $\varepsilon > 0$ and i. Taking a convergent subsequence for $\varepsilon \downarrow 0$ we may consider $\lambda_i^\varepsilon \to \lambda_i > 0$ with $x_i^\varepsilon \to x$ and note that $\langle (-y^*, 1), (x, g(x)) - (x_i, g(x_i)) \rangle \geq 0$ implies $x_i = x$ as $(x, g(x))$ is strongly exposed in epi g. As we only require to consider convex combination of length $n + 1$ (Caratheodory's theorem) we have (using $x_i^\varepsilon \in D$ a bounded set and $\lambda_i^\varepsilon \to 0$ if $x_i^\varepsilon \nrightarrow x$) we obtain

$$y^\varepsilon = \sum_{i=0}^n \lambda_i^\varepsilon x_i^\varepsilon \to \sum_{i=0}^n \lambda_i x = x \left(\sum_{i=0}^n \lambda_i \right) = x$$

contradicting $\|(y^\varepsilon, h(y^\varepsilon)) - (x, h(x))\| \geq \delta > 0$ for all $\varepsilon > 0$.

Hence the strongly exposed points of h are dense in epi h. Now the extremal points of h are contained in the convex closure of the strongly exposed points i.e. $\overline{\text{co}}\,(\text{ext epi } h) = \overline{\text{co}}\,(\exp \text{epi } g)$. Using Lemma 11 the extremal points are actually limits of strongly exposed points. As strongly exposed points are also extremal points we find that the extremal points of h are dense in epi h. For all $(x, h(x)) \in \text{ext epi } h = \text{ext co }(g)$ we have $h(x) = g(x)$. Thus h and g coincide on the dense set of exposed points of epi h (and epi g). Now apply Proposition 2 to deduce that epi $h|_D := \overline{\exp \text{epi } g}|_D$. That is for $y \in D$ we have

$$h(y) = \liminf_{\substack{(x, h(x)) \in \exp \text{epi } h \\ x \in D,\ x \to y}} g(x) \geq g(y).$$

But by construction $g(y) \geq h(y)$ so $g(y) = h(y) = \text{co } g(y)$ for $y \in D$. Apply Lemma 10 for the last observation. $\qquad \square$

Let $\bar{\eta} > 0$ be the threshold value for which Proposition 1 holds. Take $\eta > \bar{\eta}$. We wish to the Minty parametrization of a maximal monotone operator. Consider $y^*(x) = \langle x, z \rangle$ and the problem of minimizing $g(y) - \langle y, z \rangle$. The optimality conditions are

$$\begin{aligned} 0 &\in \partial f(y) + \eta(y - x^0) - z \\ &= [\partial f + \eta I](y) - (\eta x^0 + z) \\ \text{or} \qquad y &\in [\partial f + \eta I]^{-1} \left(\eta \left[x^0 + \frac{1}{\eta} z \right] \right). \end{aligned} \qquad (81)$$

As z is arbitrary we may choose $x^0 + \frac{1}{\eta} z_x = x \in \text{lev}_{x^0} f$ or $z_x = \eta(x - x^0)$. The optimality conditions then become

$$y \in [\partial f + \eta I]^{-1}(\eta x).$$

On the other hand consider the problem $y \in P_\lambda(x)$ which has optimality conditions

$$
\begin{aligned}
0 &\in \partial f(y) + \frac{1}{\lambda}(y - x) \\
\text{or} \quad y &\in \left[\partial f + \frac{1}{\lambda}I\right]^{-1}\left(\frac{1}{\lambda}x\right).
\end{aligned}
\tag{82}
$$

We may then see the correspondence under the identification $\eta = \frac{1}{\lambda}$. We may rewrite (82) as

$$
\begin{aligned}
y &\in [\partial g + (\eta - \bar\eta)I]^{-1}\left(\eta\left(x - x^0\right)\right) \\
&= [\partial g + (\eta - \bar\eta)I]^{-1}(z_x)
\end{aligned}
\tag{83}
$$

By (78) there we may take η larger if needed to ensure for all $x \in B_{\delta_\eta}(\bar x)$ and any $\bar x \in \mathrm{lev}_{x^0}f$ we have $y \in P_{\frac{1}{\eta}}(x)$ contained in $B_{\varepsilon_\eta}(\bar x) \supseteq B_{\delta_\eta}(\bar x)$ where $g(x) := f(x) + \frac{\eta}{2}\|x - x^0\|^2$ is convex on $B_\delta(\bar x) \supseteq B_{\varepsilon_\eta}(x_i)$. Assuming $\mathrm{lev}_{x^0}f$ is bounded we may extract a finite sub-cover of the open cover

$$
\left\{B_{\delta_\eta}(\bar x) \mid \bar x \in \mathrm{lev}_{x^0}f\right\}.
$$

Let

$$
\{B_{\delta_i}(x_i) \mid i = 1,\ldots,k\}
$$

be this finite cover. Then for any $x \in \mathrm{lev}_{x^0}f$ we have $x \in B_{\delta_i}(x_i)$ and hence $y \in P_{\frac{1}{\eta}}(x)$ is contained in some $B_{\varepsilon_i}(x_i) \supseteq B_{\delta_i}(x_i)$ on which g is convex. Then (83) applies so via Minty's resolvant theorem

$$
y \in T\left(\eta\left(x - x^0\right)\right) := [\partial g + (\eta - \bar\eta)I]^{-1}\left(\eta\left(x - x^0\right)\right)
\tag{84}
$$

where $T(x) := [\partial g + (\eta - \bar\eta)I]^{-1}\left(\eta\left(x - x^0\right)\right)$ is a single valued maximal monotone and nonexpansive when ∂g is maximal monotone, see [22, Theorem 12.15].

Theorem 6. *Suppose f is prox-regular and locally Lipschitz on a bounded level set $\mathrm{lev}_{x^0}f$ with $\mathrm{int}\,\mathrm{lev}_{x^0}f \neq \emptyset$. Let $g(y;x,a)$ be defined in (17) with $a \geq 0$. There exists an ath and a globally convex function $H(y;x,a)$ satisfying $g(y;x,a) \geq H(y;x,a)$ for all $y \in \mathbb{R}^n$, $a \geq a^{th}$ and $x \in \mathrm{lev}_{x^0}f$, such that $g(y;x,a)$ is the restriction to $\mathrm{lev}_{x^0}f$ of $H(y;x,a)$.*

Proof. We wish to apply Corollary 1. We take $D = \mathrm{lev}_{x^0}f$ and $g : D \to \mathbf{R}$ is Lipschitz and prox-regular on the bounded domain D. We need to show that $0 \in \partial[g - \langle z, \cdot\rangle](y)$ has a unique solution for all $y^* \in B_1(0)$ such that $(-z, 1)$ supports epi g at some $x \in D$. To this end take $x \in D$ and let $z_x = \eta(x - x^0)$ so that $g(y) - \langle y, z\rangle$ attains its minimum at y when (84) holds. But by construction we have $x \in B_{\delta_i}(x_i)$ and hence $y \in P_{\frac{1}{\eta}}(x)$ is contained in some $B_{\varepsilon_i}(x_i) \supseteq B_{\delta_i}(x_i)$ on which g is convex. Hence the operator $T(x) := [\partial g + (\eta - \bar\eta)I]^{-1}\left(\eta\left(x - x^0\right)\right)$ will have $B_{\delta_i}(x_i)$ in its domain and $B_{\varepsilon_i}(x_i)$ in its range, all contained in a region on which ∂g is locally a maximal monotone operator (as g is locally convex). Thus $y = T(x)$ is unique by [22, Theorem 12.15]. Thus $(x, g(x))$ is exposed by $(-z_x, 1)$ and Corollary 1 applies. $\square$